Text Summarization and Categorization for Scientific and Health-Related Data

A Dissertation,
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Doctor of Philosophy
in Computer Science

By

Arman Cohan, M.S.

Washington, DC
March 19, 2018

Text Summarization and Categorization for Scientific and Health-Related Data

Arman Cohan, M.S.

Dissertation, Advisor: Nazli Goharian, Ph.D.

Abstract

The increasing amount of unstructured health-related data has created a need for intelligent processing, summarizing, and categorizing these data to extract knowledge from them. My research goal in this dissertation is to develop Natural Language Processing (NLP) and Information Retrieval (IR) methods for better processing and understanding health-related textual information to promote health care and well-being of individuals.

First, I focus on scientific literature as an important source of knowledge distribution in health care. It has become a challenge for researchers to keep up with the increasing rate at which scientific findings are published. To address this problem, I propose summarization methods using citation texts and discourse structure of the papers to provide a concise representation of important contributions of the papers. I also investigate methods to address the problem of citation inaccuracy by linking the citations to their related parts in the target paper, capturing their relevant context. In addition, I raise the problem of the inadequacy of current evaluation metrics for scientific document summarization and present a superior method based on semantic relevance in evaluating the summaries.

In the second part, I focus on other significant sources of health-related information including clinical notes and social media. I investigate categorization methods to address the critical problem of medical errors which are among leading causes of

death worldwide. I demonstrate how we can effectively identify significant errors and harmful cases through medical narratives that could help prevent similar future problems. Mental health is another significant dimension of health and wellbeing that is sometimes overlooked. Suicide, the most serious challenge in mental health, accounts for approximately 1.4% of all deaths and approximately one person dies by suicide every 40 seconds. I investigate social media as a platform through which mental problems such as depression and self-harm can be investigated. I present both feature-rich and neural network methods for assessing the risk of depression, self-harm, and suicide to the individuals based on their general language expressed in social media.

INDEX WORDS: Natural Language Processing, Information Retrieval, Scientific Literature, Health-Related Text, Text Summarization, Social Media, Mental-Health

Acknowledgments

This dissertation would have not been possible without the support of many people. First and foremost, I would like to thank my advisor Nazli Goharian. Throughout my Ph.D., you have provided me with detailed and precise guidance and direction and at the same time ample freedom in research. I am grateful for your remarkable insights and support in all matters, both personal and professional. Thank you for reading countless drafts of often last minute papers and for always making time to discuss my research. I especially want to thank you for your perspective and helping me pursue and define projects with real impact.

I am also grateful to Ophir Frieder for all our discussions and his insightful advise and comments on my research and beyond. I would like to thank the rest of my amazing committee, Nathan Schneider, Elad Yom-Tov, Jimmy Lin and Calvin Newport who provided me with valuable and constructive feedback and comments on my dissertation.

Research is never done in a vacuum and I am deeply thankful to my co-authors and colleagues. I am fortunate to count many of them as friends. I would like to thank Luca Soldaini, for all the research, technical, and friendly discussions that we had. I also thank Andrew Yates who greatly inspired me in research when I first joined the Ph.D. program. I would also like to thank Medstar Institute for Innovation for amazing internship experiences and all the interesting collaborations we had. In particular, I had a great time working with Allan Fong, Ross Filice, and Raj Ratwani and attempting to solve real-world problems in health-care. I spent a summer at

Adobe Research, where I am grateful to Walter Chang, Doo Soon Kim, and Trung Bui for being supportive mentors and to Franck Dernoncourt for fruitful discussions during my abstractive summarization project. In addition, I want to thank my other amazing co-authors: Sydney Young and Sean Macavaney. I would also like to thank Pedro Fernandes, Reza Khani, and Mohammad Zaheri for being such wonderful friends and their help when I most needed it.

I would not be where I am today without the amazing encouragement, support and love from my parents Shahla and Ebrahim Cohan. I am deeply indebted to my extraordinary wife, Maryam, throughout this journey, for our many wonderful years and always standing beside me, even when more than 6,000 miles were between us.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of the problems addressed in this dissertation

In recent years, there has been an increased demand for use of health-related data obtained from a variety of sources including scientific literature, medical reports and notes, and social media.

In this dissertation, I argue that Natural Language Processing and Information Retrieval can help us in addressing some of the real-world challenges in the health-care. In particular, I show how we can help doctors, patients and scientists through improved and more intelligent methods for: summarizing scientific articles and distributing knowledge; analyzing textual reports of medical errors; improving education of medical students; and identifying at-risk individuals in social media.

The rapid growth of scientific literature has made it difficult for researchers to find an overview of the latest developments in their respective fields. The existence of surveys in various fields show that such information is desirable, yet procuring such surveys, given the fast publication rates, requires painstaking work. A recent study showing that global scientific output is doubling every nine years [25] further demonstrates the significance of this challenge. Automatic summarization of scientific literature is one way to address this challenge. Recently citation-based summarization approaches have been proposed to address the shortcomings of abstract-as-summary approach [216, 219]. In these methods, a set of citation texts (i.e., textual spans

surrounding a citation explaining the referenced work) is used to capture the contributions of a target paper. While these methods have shown to be effective, citation texts which are describing a specific contribution of a referenced paper can be inaccurate in conveying the exact information from the referenced paper. The inaccuracy of citations can be attributed to the fact that they are written by different authors. The citing authors might misunderstand some points of the referenced paper, they might ascribe contributions to the reference paper which are not existent, and they might only mention results without discussing the assumptions, datasets, or experimental conditions under which those results were obtained. These problems can have severe negative outcomes. For example in life sciences and biomedicine, findings of papers can directly or indirectly impact human lives and inaccurate citations might result in an inaccurate summary and have adverse future consequences. Therefore, there is a need for improving the way citations are quoting the referenced paper. If this limitation of citations is addressed, they can be utilized for summarizing the key contributions of a given referenced paper.

Similar to most tasks in information processing, accurate evaluation of automatic summarization systems is an important problem. Traditional evaluation of summarization involves direct human assessment of different quality metrics through predesigned questionnaires. However, conducting such evaluation is expensive, and the results are not reproducible or in some cases not reliable. Automated evaluation metrics address this challenge by methods quantifying the quality of a system generated summary against a set of gold standard summaries. While researchers have investigated summarization evaluation in the general domain, there is lack of evidence for effectiveness of such evaluation metrics in the summarization of scientific papers.

In addition to the rapid growth of biomedical literature, there is an increasing demand for use of electronic health records and clinical texts, for reasons such as

improving health care, public health surveillance, quality measures, and improving medical education. Preventable medical errors have been shown to be a major cause of injury and death in the United States [85, 267]. In many standard clinical workflows in hospitals and health centers, residents first examine cases and write a preliminary report that reflects their interpretation of the case. This initial report is then reviewed by an attending doctor who also reviews the case and revises the initial report in case of any misinterpretations or errors. The edited report is served as the official report for that case. While most of the revisions are due to different reporting styles of the resident and the attending, in some cases, the final revision reflects existence of errors in the initial report. In these cases, there are critical discrepancies between the two reports that imply the resident has made an initial misinterpretation, misdiagnoses, or wrong reporting. Addressing these situations directly affects patient care and the resident's education. The large volume of medical reports everyday makes it difficult to manually distinguish significant discrepancies from those that are merely due to reporting styles. To identify sources of common preventable errors, healthcare centers have started utilizing reporting systems to log the events occurring to the patients at the healthcare centers. These reports are usually natural language narratives describing the timeline of the patient. While these reports are useful in case by case basis, manual large-scale identification of harmful events to the patients or common sources of such problems are challenging. Automated effective categorization of the these reports and the severity of harm associated with them can greatly benefit healthcare systems.

Mental health is another dimension of health and wellbeing that is sometimes taken for granted. This is while mental health, suicide and its prevention remain major challenges in public health care. Suicide is one of the leading causes of death [186]. Each year 43,000 Americans die by suicide, on average there are 117 suicides

per day, and about 500,000 people visit hospital for injuries due to self-harm [8, 136, 186]. Each suicide case has major consequences on the physical and emotional well-being of families and on societies in general [231, 250]. Therefore, identifying individuals with depression or at risk of suicide and providing them with sufficient support remains an important problem [239]. Due to the stigma often associated with mental-health issues, many individuals tend to express their problems in an anonymous or pseudo-anonymous fashion through social media. Hence, social media has become a major platform through which mental-health issues and problems are expressed and discussed. It would be desirable to utilize data in social media to identify users that are at risk of depression or suicide and provide them with the help and resources that they need.

1.2 Contributions and Outline of This Thesis

In this thesis, my research goal is to address the real-world challenges discussed above. Specifically, my research will substantiate the following hypotheses:

H1 Improving scientific document summarization:

H1.1 Citation texts are not always accurate and thus adding context from the reference paper improves their accuracy.

H1.2 Citation contexts and scientific discourse structure can be leveraged to improve scientific document summarization.

H1.3 Existing summarization evaluation metrics are not adequate for scientific document summarization. An evaluation metric that considers similarity beyond lexical overlaps is superior.

To address the problem of citation inaccuracy, I propose methods for linking citation texts with their relevant parts in the referenced paper. These methods are presented in first part of Chapter 2. I argue that incorporation of citation contexts along with the discourse structure can improve the citation-based summarization approaches. There are two prominent approaches towards summarization: (*i*) extractive summarization where the summary is generated by copying important textual spans from the input document; (*ii*) abstractive summarization where the summary is generated from scratch and the summary might include words or phrases that are not in the input document [46, 228]. Citation-based summarization methods suffer from the cold start problem, where it is not possible to obtain a good quality summary for the newly published documents with no or only a few citations. As an alternative method, in Section 2.3, I propose an abstractive summarization method where the input is only the document and the summary is generated abstractively using the discourse structure of the document. Finally, I challenge the adequacy of current evaluation methods for summarization in the scientific domain and propose a method in Section 2.4 for improved summarization evaluation.

H2 Text categorization applications in the health-related domain:

H2.1 The differences between the initial and final versions of medical reports can be differentiated into substantive and stylistic through carefully designed features.

H2.2 A neural attention model can be used to categorize patient reports and assess the severity of patient harm in these reports.

H2.3 With Natural Language Processing methods, we can identify users at risk of depression or self-harm through social media.

Text categorization and classification is a fundamental task in understanding, mining, and analyzing medical text and it can benefit applications that improve healthcare in general. In Chapter 3, my research goal is to develop NLP methods to address some of the real-world problems regarding healthcare and wellbeing of individuals. I discuss methods to differentiate between errors and stylistic problems in initial and final versions of medical reports (Section 3.2). I will then present a neural attention model that can effectively categorize patient reports to their respective harm categories (Section 3.3). Later in Section 3.4, I will propose methods to identify users that are at risk of self-harm or suicide in the specialized mental-health online forums. I will then switch focus from specialized mental-health forums to to general forums in Section 3.4.4, and propose methods to assess depression risk in the users. In the same section, I will also discuss our data collection methods that enable other researchers to further explore mental-health challenges through social media.

This dissertation is based on my following research publications: [48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 221, 273]. I have been the main contributor to the ideas, implementations, and writings of all my referenced publications.[1]

1.3 Organization

The remaining chapters in this dissertation proposal are organized as follows. In Chapter 2, I address *H1* by *(i)* proposing methods for improving the accuracy of the citation texts through contextualization; *(ii)* utilizing citations for enhancing summarization in the scientific domain; *(iii)* proposing a discourse-aware abstractive method for summarization of scientific papers; and *(iv)* describing a method to improve summarization evaluation.

[1]For the EMNLP 2017 paper [273], Andrew Yates and I equally contributed to the work.

Chapter 3 addresses *H2* by *(i)* describing methods for differentiating errors from stylistic discrepancies in clinical reports; *(ii)* explaining how we can identify and categorize preventable harmful events in medical reports; and *(iii)* outlining approaches for improved self-harm and depression risk assessment in social media.

Finally, Chapter 4 concludes my dissertation.

Chapter 2

Scientific Document Summarization

2.1 Introduction

In scientific literature, related work is often referenced along with a short textual description regarding that work which we call citation text. Citation texts usually highlight certain contributions of the referenced paper and a set of citation texts to a reference paper can provide useful information about that paper. Therefore, citation texts have been previously used to enhance many downstream tasks in IR/NLP such as search and summarization [48, 216, 224].

At the same time, keeping up with the new scientific developments has become challenging due to the increasing rate of publications [233]. Summarizing the key contributions and findings of papers can help researchers to find out about new ideas and findings in scientific fields. Scientific papers are accomponied with an abstract that usually inculdes a summary of the paper written by the same authors. Although abstracts provide an overview of the papers, occasionally some key information are missing, or the contributions stated in abstracts are overstated. These problems have inspired another type of scientific summarization using citation texts [216]. This type of summaries draw key contributions from a target paper using a set of citation texts.

While useful, citation texts might lack the appropriate context from the reference article [48, 80, 257]. For example, details of the methods, assumptions or conditions for the obtained results are often not mentioned. Furthermore, in many cases the

citing author might misunderstand or misquote the referenced paper and ascribe contributions that are not intended. Hence, sometimes the citation text is not sufficiently informative or in other cases, even inaccurate [233]. This problem is more serious in life sciences where accurate dissemination of knowledge has direct impact on human lives.

In this chapter, I first describe an extractive summarization method which utilizes citations and article discourse structure for summarizing key findings of the paper. In the second part, I present an alternative abstractive approach for generating scientific summaries. Finally, I conclude this chapter with a discussion of evaluation metrics for summarization in the scientific domain and propose a new evaluation metric which improves over existing evaluation methods.

2.2 Extractive Summarization using Citation Contextualization and Scientific Discourse

2.2.1 Introduction

Abstracts are a basic form of scientific summary written by the author of the paper, explaining contributions and points of the paper. While abstracts provide an overview of the paper, they do not necessarily convey all the important contributions and impacts of the paper [90]: (*i*) The authors might ascribe contributions to their papers that are not existent. (*ii*) some important contributions might not be included in the abstract; (*iii*) the contributions stated in the abstract do not convey the article's impact over time and comparisons with future related work are not possible through abstracts; (*iv*) abstracts usually provide a very broad view of the papers and they may not be detailed enough for people seeking detailed contributions; (*v*) The content distribution in the abstracts are not evenly drawn from different sections of the papers [10]. These problems have inspired another type of scientific summary which is obtained by utilizing a set of citations referencing the original paper [216, 219]. Each citation is often accompanied by a short description explaining the ideas, methods, results, or findings of the cited work. This short description is called citation text or citance [184]. Therefore, a set of citation texts by different papers can provide an overview of the main ideas, methods and contributions of the cited paper, and thus, can form a summary of the referenced paper. These community based summaries capture the important contributions of the paper, view the article from multiple aspects, and reflect the impact of the article to the community.

At the same time, there are multiple problems associated with citation texts. They are written by different authors so they may be biased toward another work. The citation texts lack the context in terms of the details of the methods, the data,

Reference Article

(Voorhoeve et al., 2006): "These miRNAs could neutralize p53-mediated CDK inhibition, possibly through direct inhibition of the expression of the tumor suppressor LATS2."

Citing Articles

(Kloosterman and Plasterk, 2008): "In a genetic screen, miR-372 and miR-373 were found to allow proliferation of primary human cells (Voorhoeve et al., 2006)."
(Okada et al., 2011): "Two oncogenic miRNAs, miR-372 and miR-373, directly inhibit the expression of Lats2, thereby allowing tumorigenic growth in the presence of p53 (Voorhoeve et al., 2006)."

Figure 2.1: Example of epistemic value drift. The claims that Voorhoeve et al. (2006) state as possibilities, becomes fact in later citations (Okada et al., 2011; Kloosterman and Plasterk, 2008).

assumptions, and results. More importantly, the points and claims by the original paper might be misunderstood by the citing authors; certain contributions might be ascribed to the cited work that are not on par with the original author's intent. Another serious problem is the modification of the epistemic value of claims, which states that many claims by the original author might be stated as facts in the future citations [81]. An example of this is shown in Figure 2.1. As illustrated, while the original authors write on some possibilities, later the citing authors state them as known facts. These problems are even more serious in biomedical domain where slight misrepresentations of the specific findings about treatments, diagnosis, and medications, could directly affect human lives.

One way to address such problems is to consider the citations in their context from the reference article. Therefore, citation texts should be linked to the specific parts in the reference paper that correctly reflect them. We call this "citation contex-

tualization". Citation contextualization is a challenging task due to the terminology variations between the citing and cited author's language usage.

Scientific papers have the unique characteristic of following a specific discourse structure. For example, a typical scientific discourse structure follows this form: problem and motivation, methods, experiments, results, and implications. The rhetorical status of a citation provides additional useful information that can be used in applications such as information extraction, retrieval, and summarization [258]. Each citation text could refer to specific discourse facets of the referenced paper. For example one citation could be about the main method of the referenced paper while the other one could mention their results. Identifying these discourse facets has distinct values for scientific document summarization; it allows creating more coherent summaries and diversifying the points included in the generated scientific summaries.

Scientific document summarization is recently further motivated by TAC[1] 2014 summarization track, and the 2016 computational linguistics summarization shared task [121]. Following these works and motivated by the challenges mentioned above, we propose an extractive approach for scientific document summarization based on citations. Our framework improves the shortcomings of existing citation-based summarization methods such as [216]. We particularly first address the problem of citation inaccuracy in conveying information from the referenced paper. To address this problem, we add the relevant context from the referenced paper to the citation. We then utilize the article's discourse structure to group similar content together and finally we select important content for the summary. In particular, our method consists of the following steps:

- *Contextualizing citation texts*: Citations are not always accurate in conveying the information in the reference paper. One approach to address this problem

[1]Text Analysis Conference, http://tac.nist.gov/2014/BiomedSumm/

is to link citation texts with their relevant parts in the referenced paper. This relevant parts provide context for this citation text and we call this process citation contextualization. We propose several approaches for contextualizing citations. Finding the exact reference context for the citations is challenging due to discourse variation and terminology differences between the citing and the referenced authors. Therefore, traditional Information Retrieval (IR) methods are inadequate for finding the relevant contexts. We propose to address this challenge by three approaches: *(i)* query reformulations, *(ii)* utilizing word embeddings [13], and domain-specific knowledge and *(iii)* supervised classification. In these models, our goal is to address the terminology variation problem between the citing and cited authors.

- *Discourse structure*: Scientific papers usually follow a standard discourse structure where the authors first introduce the problem, then they talk about the scope and methodology, experimental setup, results, discussions and finally conclusions [251]. A good summary should capture information from all these different discourse facets. Hence, after extracting the context of the citation texts, we group them into different discourse facets of the article. We use a linear classifier with variety of features for classifying the citations.

- *Summarization*: We propose two approaches for summarizing the papers. Both approaches are based on summarization through the scientific community where the main points of a paper are captured by a set of given citations. Our approach extends the previous works on citation-based summarization [216, 217, 218] by incorporating the reference context to address the inaccuracy problem associated with the citation texts. After extracting the citation contexts from the

reference paper, we group them into different discourse facets. Then using the most representative sentences in each group, we generate the final summary.

In particular our contributions are summarized as follows: (*i*) Methods for contextualizing the citation texts from the reference article. (*ii*) Identifying the discourse facets of the citation contexts. (*iii*) A scientific document summarization approach utilizing citation contexts and the scientific discourse structure. (*iv*) Extensive evaluation on two scientific domains.

2.2.2 Related work

2.2.2.1 Citation text analysis

Citations play an integral role in the scientific development. They help disseminate the new findings and they allow new works to be grounded on previous efforts [111]. While there is a large body of related work on analysis of citation networks, instead of link analysis, we focus on textual aspects of the citations. To better utilize the citations, researchers have explored ways to extract citation texts, which are short textual parts describing some aspects of the cited work. Examples of the proposed approaches for extracting the citation texts include jointly modeling the link information and the citation texts [137], supervised Markov Random Fields classifiers [217], and sequence labeling with segment classification [3]. These approaches focus on finding the sentences or textual spans in the citing article that explain some aspects of the cited work. In this work, we assume that citation texts are already obtained either manually or by using one of these works. Given the citation texts, we instead focus on contextualizing these citation texts using the reference; we find the text spans in the reference article that most closely reflect the citation text.

There exists some related work on further analyzing the citations for finding their function or rhetorical status [4, 97, 111, 258]. In these works, the authors tried to identify the reasons behind citations which can be a statement of weakness, contrast or comparison, usage or compatibility, or a neutral category. They proposed a classification framework based on lexically and linguistically inspired features for classifying citation functions. The distribution of citations within the structure of scientific papers have been also studied [18]. The authors of [36] have investigated the problem of measuring the intensity of the citations in scientific papers and in [37], the authors proposed using the discourse facets for scientific article recommendation. Recently, a framework for understanding citation function has been proposed [132] which unifies all the previous efforts in terms of definition of citation functions. While citation function can provide additional information for summarization, in this work we do not utilize these information. Instead, we utilize the discourse facet of the citation contexts in a reference paper.

2.2.2.2 Citation contextualization

More recently, there has been some efforts in contextualizing citations from the reference. In particular, TAC 2014 summarization track,[2] and the CL-SciSumm 2016 shared task on computational linguistic summarization [121] have released datasets to promote research for citation contextualization. The former is more domain specific, focusing on biomedical scientific literature, while the latter is in a more general domain consisting of publications in computational linguistics. To our knowledge, there is no overview paper on TAC. We briefly discuss the successful approaches in CL-SciSumm 2016. The authors of [32] used an SVM-rank approach with features

[2] http://tac.nist.gov/2014/BiomedSumm/

such as tf-idf[3] cosine similarity, position of the reference sentence, section position, and named entity features. In another approach [153], the authors used an SVM classifier with sentence similarity and lexicon based features. The authors of [193] proposed a hybrid model based on tf-idf similarity and a single layer neural network that scores the relevant reference texts above the irrelevant ones. Finally, in the work by [146], the authors proposed the use of TextSentenceRank algorithm which is an enhanced version of the TextRank algorithm for ranking keywords in the documents. Here, we specifically focus on the problem of terminology variation between the citing and cited authors. We propose approaches that address this problem. Our proposed approaches are based on query reformulations, word embeddings, and domain-specific knowledge.

2.2.2.3 Text summarization

Document summarization has been an active research area in NLP in recent decades; there is a rich literature on text summarization. Approaches towards summarization can be divided into the following categories: *(i)* topic modeling based [34, 100, 248, 262]: In these approaches, the content or topical distribution of the final summary is estimated using a probabilistic framework. (*ii*) solving an optimization problem [17, 47, 88]: these approaches cast the summarization problem as an optimization problem where an objective function needs to be optimized with respect to some constraints. (*iii*) supervised models [38, 62, 197], where selection of sentences in the summary are learned using a supervised framework. (*iv*) graph based [92, 171, 204]: these approaches seek to find the most central sentences in a document's graph where sentences are nodes and edges are similarities. (*v*) Heuristic based [33, 107, 156]: these works approach the summarization problem by greedy selection of the content.

[3]Term Frequency - Inverted Document Frequency.

(*vi*) Neural networks: More recently, there has been some efforts on utilizing neural networks and sequence-to-sequence models [253] for generating summaries of short texts and sentences [45, 227]. Most of these works have focused on general domain summarization and news articles. Scientific articles are much different than news articles in elements such as length, language, complexity and structure [256].

One of the first works in scientific article summarization is done by [256] where the authors trained a supervised Naive Bayes classifier to select informative content for the summary. Later, the impact of citations to generate scientific summaries was realized [90]. In the work by [218], the authors proposed an approach for citation-based summarization based on a clustering approach, while in [2] and [124], the focused on producing coherent scientific summaries. We argue that citation texts by themselves are not always accurate and they lack the context of the cited paper. Therefore, if we only use the citation texts for scientific document summarization, the resulting summary would potentially suffer from the same problems, and it might not accurately reflect the claims made in the original paper. We address this problem by leveraging the citation contexts from the reference paper. We also utilize the inherent discourse structure of the scientific documents to capture the important content from all sections of the paper.

We present a comprehensive framework for scientific document summarization which utilizes and builds upon our earlier efforts [48, 51, 52]. We propose new approaches for citation contextualization. We further extend our experiments on an additional dataset (CL-SciSumm 2016) and evaluate our approaches on both TAC and CL-SciSumm datasets, providing detailed analysis.

2.2.3 Methodology

Our proposed method is a pipeline for summarizing scientific papers. It consists of the following steps:

1. citation contextualization (extracting the relevant context from the reference paper)
2. identifying the discourse facet of the extracted context
3. summarization

We first explain our proposed methods for contextualization, we then describe our approach for identifying discourse facets of the citation contexts, and finally we outline our summarization approach.

2.2.3.1 Citation contextualization

Citation contextualization refers to extracting the relevant context from the reference article for a given citation text. We propose the following three approaches for this problem: (*i*) Query reformulation, (*ii*) Word embeddings and domain knowledge, and (*iii*) Supervised classification.

Query reformulation (QR). We cast the contextualization problem as an Information Retrieval (IR) task. We first extract textual spans from the reference article and index them using an IR model. The textual spans are of granularity of sentences. In order to capture longer contexts (those consisting of multiple consecutive sentences), we also index sentence n-grams. That is, we index each n consecutive sentences as a separate text span.[4] After constructing the index, we consider the citation text as

[4]we indexed up to 3 consecutive sentences in our experiments.

the query, and we seek to find the relevant context from the indexed spans. Since the citation texts are often longer than usual queries in standard IR tasks, we apply query reformulation methods on the citation to better retrieve the related context. We utilize both general and domain-specific query reformulations for this purpose. We first remove the citation markers (author names and year, and numbered citations) from the citations, as they do not appear in the reference text and hence are not helpful. We design several regular expressions to capture these names.

Since the citation texts are usually more verbose than standard queries, there might be many uninformative terms in them that do not contribute in finding the correct context. Hence, we apply query reduction methods to only retain the important concepts in the citation. After removing the stop words from the citation, we further experiment with the following three query reduction methods:

1. Noun phrases (QR-NP). Citation texts are usually linguistically well-formed, as they are extracted from scientific papers. This allows us to apply a variety of linguistic tagging and chunking methods to the query to capture the informative phrases. Previous works have shown that noun phrases are good representation of informative concepts in the query [12, 119, 120]. We thus extract noun phrases from the citation text and omit all other terms. We use Stanford CoreNLP [168] for extracting the noun phrases.

2. Key concepts (QR-KW). Key concepts or keywords are single or multi-word expressions that are informative in finding the relevant context. We use the Inverted Document Frequency (IDF) [242] measure to find the key concepts. The terms that are prevalent throughout all the text spans do not provide much information in retrieval. IDF values help capturing the terms and concepts that are more specific. For key concept extraction, we limit the IDF values between

some threshold that can be tuned according to the dataset.[5] We consider phrases of up to three terms.

3. Ontology (QR-Domain). Domain-specific ontologies are expert curated lexicons that contain domain-specific concepts. In this reformulation method, we use an ontology to only keep important (domain-specific) concepts in the query. Since the TAC dataset is in the biomedical domain, we use the UMLS [24] thesaurus which is a comprehensive ontology of biomedical concepts. We specifically use the SNOMED CT [240] subset of UMLS.

As explained in Section this section, the indexing approach also contains consecutive sentences. Therefore, our retrieval approach can find text spans that have overlaps with each other. Furthermore, retrieving multiple spans from around the same location in the text signals the importance of that specific location. We apply a reranking and merging method to the retrieved spans to remove shared spans and better rank the more relevant context. We merge the two overlapping spans if the retrieval score of the larger span is higher than the smaller span. We also evaluated other query reformulation methods such as Pseudo Relevance Feedback [31]; however, they performed worse than the baseline and thus we do not discuss them further.

An information retrieval model for contextualization based on embeddings and domain ontologies. We explained how we can leverage query reformulation methods to modify the citation to make it more expressive. Instead of modifying the query, we can modify the retrieval model to directly account for terminology variations and paraphrasing between the citing and the cited authors. Specifically, we propose to

[5]We empirically set this threshold to 1.9 and 2.2 for the TAC and CL-SciSumm datasets, respectively.

achieve this using an information retrieval model based on word embeddings and domain-specific knowledge.

Embeddings. Word embeddings or distributed representations of words are mapping of words to dense vectors according to a distributional space, with the goal that similar words will be located close to each other [14]. We extend the Language Modeling (LM) for information retrieval model [213] by utilizing word embeddings to account for terminology variations. Given a citation text (query) q, and a reference span (document) d, the LM scores d based on the probability that d has generated q ($p(d|q)$). Using standard simplifying assumptions of term independence and uniform document prior, we have:

$$p(d|q) \propto p(q|d) = \prod_{i=1}^{n} p(q_i|d) \tag{2.1}$$

where q_i $(i = 1, ..., n)$ are the terms in the query. In LM with Dirichlet Smoothing [277], $p(q_i|d)$ is calculated using a smoothed maximum likelihood estimate:

$$p(q_i|d) = \frac{f(q_i, d) + \mu\, p(q_i|C)}{\sum_{w \in V} f(w, d) + \mu} \tag{2.2}$$

where f is the frequency function, $p(q_i|C)$ shows the background probability of term q_i in collection C, V is the entire vocabulary, and μ is the Dirichlet parameter.

Our model extends the above formulation (Eq. 2.2) by using word embeddings. In particular we estimate the probability $p(q_i|d)$ according to the following equation:

$$p(q_i|d) = \frac{\sum_{d_j \in d} s(q_i, d_j) + \mu\, p(q_i|C)}{\sum_{w \in V} \sum_{d_j \in d} s(w, d_j) + \mu} \tag{2.3}$$

where d_j are terms in the document d, and s is a function that captures the similarity between the terms and is defined as:

Table 2.1: Example of similarity values between terms.

word 1	word 2	Similarity
marker	mint	0.11
notebook	sky	0.07
capture	promotion	0.12
blue	sky	0.31
produce	make	0.43

The table shows an example of similarity values between terms according to the dot product of their corresponding embeddings. Pre-trained Word2Vec model on Google News corpus is used for embeddings. The top part of the table shows pairs of random words, while the bottom part shows similarity values for pairs of related words.

$$s(q_i, d_j) = \begin{cases} \phi\big(e(q_i), e(d_j)\big), & \text{if } e(q_i).e(d_j) > \tau \\ 0, & \text{otherwise} \end{cases} \tag{2.4}$$

where $e(q_i)$ shows the unit vector corresponding to the embedding of word q_i, τ is a threshold, and ϕ is a transformation function. Below we explain the role of parameter τ and the transformation function ϕ.

Word embeddings can capture the similarity values of words according to some distance function. Most embedding methods represent the distance in the distributional semantics space. Therefore, similarities between two words q_i and d_j can be captured using the dot product of their corresponding embeddings (i.e. $e(q_i).e(d_j)$). While high values of this product suggest syntactic and semantic relatedness between the two terms [113, 172, 209], many unrelated words have non-zero dot products (an example is shown in Table 2.1). Therefore, considering them in the retrieval model introduces noise and hurts the performance. We address this issue by first considering

a threshold τ below which all similarity values are squashed to zero. This ensures that only highly relevant terms contribute to the retrieval model. To identify an appropriate value for τ, we select a random set of words from the embedding model and calculate the average and standard deviation of point-wise absolute values of similarities between the pairs of terms from these samples. We then set τ to be two standard deviations larger than the average similarities, to only consider very high similarity values. We also observe that for high similarity values between the terms, the values are not discriminative enough between more or less related words. This is illustrated in Figure 2.2 where we can see that the most similar terms to the given term are not very discriminative. In other words, the similarity values decline slowly as moving away from top similar words. We instead want only very top similar words to contribute to the retrieval score. Therefore, we transform the similarity values according to a *logit* function (equation 2.5) to dampen the effect of less similar words (see Figure 2.2):

$$\phi(x) = \log(\frac{x}{1-x}) \tag{2.5}$$

While any approach for training the word embeddings could be used, we use the Word2Vec [150] method, which has proven effective in several word similarity tasks. We train Word2Vec on the recent dump of Wikipedia.[6] Since the TAC dataset is in biomedical domain, we also train embeddings on a domain-specific collection; we use the TREC Genomics collections, 2004 and 2006 [112] which together consist of 1.45 billion tokens.

Incorporating domain knowledge Word embedding models learn the relationship between terms by being trained on a large corpus. They are based on the distribu-

[6] `https://dumps.wikimedia.org/enwiki/`

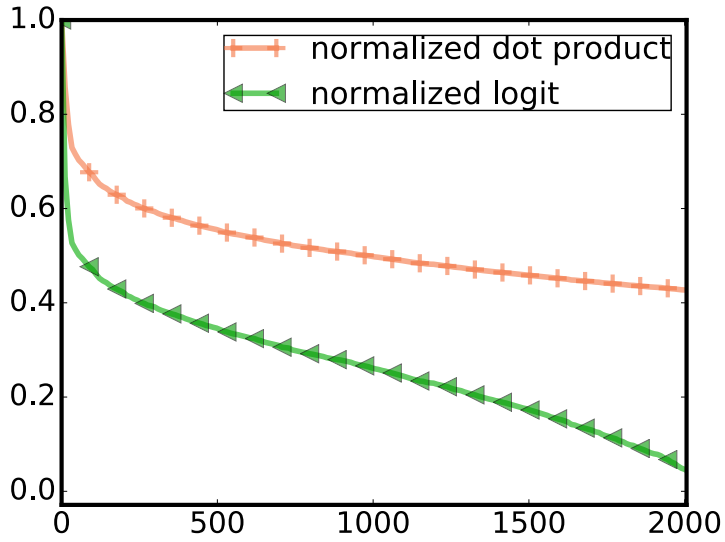


Figure 2.2: Normalized similarity values in an embedding space. The x axis is the word indexes and the y axis is the similarity values. The orange line with + markers shows the original similarity values, while the green line with triangle markers shows the transformed values using the logit function. The logit function, dampens the similarity values of less similar words.

tional hypothesis [109] which states that similar words appear in similar contexts. While these models have been very successful in capturing semantic relatedness, recent related works have shown that domain ontologies and expert curated lexicons may contain information that are not captured by embeddings [94, 113, 181]; hence, we account for the domain knowledge according to the following.

- Retrofitting embeddings: In this method, we apply a post-processing step called retrofitting [94] to the word embeddings used in the model. Retrofitting optimizes an objective function that is based on relationships between words in a lexicon; it intuitively pulls closer the words that are related to each other and

pushes farther the words that are not related to each other according to a given ontology. For the ontology, since TAC data is in biomedical domain we use two domain-specific ontologies, MESH[7] [158] and Protein Ontology (PRO).[8]. UMLS is another widely used ontology in the medical domain that could be used and MESH is one of its subsets. UMLS is very broad and contains entries even for general-domain words. Here, we need a more focused ontology to only capture biomedical specific relations. Hence, we opted for MESH and the PRO ongtologies.

For the CL-SciSumm data, since it is less domain-specific, we use the WordNet lexicon [174].

- Interpolating in the LM: In this method, instead of modifying the word vectors, we incorporate the domain knowledge directly in the retrieval model. We do so by interpolation of two following probability estimates:

$$p(q_i|d) = \lambda p_1(q_i|d) + (1-\lambda)p_2(q_i|d) \tag{2.6}$$

where p_1 is estimated using Eq. 2.3 and p_2 is a similar model that counts in the *is-synonym* relations (is-syn) in calculating similarities. Its formulation is exactly like Eq. 2.3 except it replaces the function s with the following function:

$$s_2(q_i, d_j) = \begin{cases} 1, & \text{if } q_i = d_j \\ \gamma, & \text{if } q_i \text{ is-syn } d_j \\ 0, & \text{o.w.} \end{cases} \tag{2.7}$$

[7]MEdical Subject Headings
[8]http://pir.georgetown.edu/pro/

This function is essentially partially counting the synonyms in calculation of the probability estimate $p(q_i|d)$ by the amount of γ. We empirically set the value of γ. Word embedding based methods are abbreviated by WE in the results.

Supervised classification. The two previous context retrieval models are unsupervised with respect to the contextualization task and as such, do not take advantage of the already labeled data. The CL-SciSumm dataset includes separate training and testing sets which allow us to also investigate supervised approaches. We propose a feature-rich classifier to find the correct context for each given citation. Our approach aims to capture the semantic relatedness between a given citation text and a candidate context sentences[9]. We consider all the sentences in the paper (except for the references) as candidates. We specifically utilize the following features to capture this relatedness:

- Word match: counts the number of identical words between the source citation text and the candidate reference context normalized by length.
- Fuzzy word match: same as above, with the difference that we use character n-grams to capture partial matches between the words.
- Embedding-based alignment: measures the similarity between the source and target sentences using word embedding alignment. Specifically for the two sentences S_1 and S_2, the following function f scores the sentences based on their similarity:

$$f(S_1, S_2) = \frac{\sum_{w \in S_1} \max_{v \in S_2} s(w, v)}{|S_1|} \tag{2.8}$$

[9]In CL-SciSumm dataset the gold context is of granularity of sentences.

where s is a similarity function according to the equation 2.4. Intuitively, f captures the similarity between the two sentences without only relying on lexical overlaps; it takes into account the similarity values between the terms.

- Distance between average of embeddings: measure the similarity between the two sentences by dot product of the average of their constituent word vectors.
- BM25 similarity score [225] between the citation text and the candidate text span.
- Tf-idf and count vectorized similarities: dot product between the sparse tf-idf weighted or count weighted vectors associated with the source citation and target text span.
- Character n-gram Tf-idf and count vectorized similarities: same as above, except that we used 3-gram characters to allow partial word matches.

We train a standard linear classifier (e.g. Logistic Regression) using these features to identify the correct context for a given citation text.

2.2.3.2 Generating the summary

After extracting reference contexts for the citations as described in Section 2.2.3.1, we generate a summary of the reference paper. Our goal is to create a summary that contains information from different discourse facets of the paper. This helps not only in diversifying the content in the summary, but also in creating a more coherent summary. We present the following methods for grouping citation-contexts:

Grouping the citation-contexts. After identifying the context for each citation, we use them to form the summary. To capture various important aspects of the reference

article, we form groups of citation-contexts that are about the same topic. We use the following two approaches for forming these groups:

Community detection. We want to find diverse key aspects of the reference article. We form the graph of extracted reference spans in which nodes are sentences and edges are similarity between sentences. As for the similarity function, we use cosine similarity between tf-idf vectors of the sentences. Similar to [216], we want to find subgraphs or communities whose intra-connectivity is high but inter-connectivity is low. Such quality is captured by the modularity measure of the graph [189, 190]. Graph modularity quantifies the denseness of the subgraphs in comparison with denseness of the graph of randomly distributed edges and is defined as follows:

$$Q = \frac{1}{2m} \sum_{vw} \big[A_{vw} - \frac{k_v \times k_w}{2m}\big] \delta(c_v, c_w)$$

Where A_{vw} is the weight of the edge (v, w); k_v is the degree of the vertex v; c_v is the community of vertex v; δ is the Kronecker's delta function and $m = \sum_{vw} A_{vw}$ is the normalization factor.

While the general problem of precise partitioning of the graph into highly dense communities that optimizes the modularity is computationally prohibitive [27], many heuristic algorithms have been proposed with reasonable results. To extract communities from the graph of reference spans, we use the algorithm proposed in [23] which is a simple yet accurate and efficient community detection algorithm. Specifically, communities are built in a hierarchical fashion. At first, each node belongs to a separate community. Then nodes are assigned to new communities if there is a positive gain in modularity. This process is applied iteratively until no further improvement in modularity is possible.

Table 2.2: Features for identifying discourse facets.

Feature Name
Citation Text
Extracted Reference Context
Verb Features
Ralative Section Position

Discourse model. The organization of scientific papers usually follows a standardized discourse pattern, where the authors first describe the problem or motivation, then they talk about their methods, then the results, and finally discussion and implications [251]. Our goal is to capture the important content from all sections of the paper; therefore, after extracting the citation contexts, we identify the associated discourse facet for each of the citation contexts retrieved from the previous step. Each citation context refers to some specific discourse facets of the reference document. To identify the correct discourse facets, we train a simple supervised model with features listed in Table 2.2. Essentially, we use the citation text and the extracted reference context represented by character n-grams, the verbs in the context sentence, and the relative position of the retrieved context in the paper as features for the classifier. While the textual features (citation and its context) were the most helpful, we empirically observed slight improvements by incorporating the verb and section position features. We train the model using an SVM classifier [266]. For the textual features, we transform them using character n-grams to allow fuzzy matching between the terms. Both TAC and CL-SciSumm datasets include annotated data for discourse facets which make training the supervised models possible.

Ranking model. To identify the most representative sentences of each group, we require a measure of importance of sentences. We consider the sentences in a group as a graph and rank nodes based on their importance. In particular, we consider sentences in each group as nodes and their similarities as weighted edges in a graph. An important node is a node that has many connections with other nodes. There are various ways of measuring centrality of nodes such as nodes' degree, betweenness, closeness and eigenvectors. Here, we opt for eigenvectors and we find the most central sentences in each group by using the "power method" [92] which is a random-walk based method by iteratively updates the eigenvector until convergence. It works by iteratively updating the score of each sentence according to its centrality (total weight of incoming edges) and the centrality of its neighbors. After ranking the sentences in each group according to their centrality score, we select sentences for the final summary. We use the following methods for creating the final summary:

- Iterative. This method simply iterates over the discourse facets and selects the top representative sentence from each group until the summary length threshold is met.

- Greedy. The iterative approach could result in similar sentences ending up in the summary; this results in redundant information and potential exclusion of other important aspects of the paper from the summary. To address this potential problem, we use a heuristic that accounts for both the informativeness of candidate sentence and their novelty with respect to what is already included in the summary. Maximal Marginal Relevance [33] is one such heuristic that has these properties. It is based on the linear interpolation of the informativeness and the novelty of the sentences.

2.2.4 Experiments

2.2.4.1 Data

We conducted our experiments on two scientific document summarization datasets. The first dataset is the TAC 2014 scientific document summarization dataset.[10] The TAC benchmark is in biomedical domain and is publicly available upon request from NIST.[11] The second dataset is the 2016 CL-SciSumm dataset [121] which is available on a public repository[12] and contains scientific articles from the computational linguistics domain. To our knowledge, these two are the only datasets on scientific document summarization.

The TAC dataset only has one training set consisting of 20 topics. There is one reference article in each topic and another set of articles citing the reference. For each topic, 4 annotators have identified the relevant contexts, the correct discourse facet, and they have written a summary. The documents are provided as plain text files and there are no predefined sentence boundaries and sections. On the other hand, the CL-SciSumm data contain separate train, development, and test sets with 30 topics in total. Similar to TAC, each topic consists of reference and a set of citing articles but in the computational linguistics domain. The articles are in xml format with known sentence boundaries and sections. Another distinction is that topics in the CL-SciSumm data are annotated by one annotator at a time. The full statistics of the datasets is illustrated in Table 2.3. The distribution of the discourse facets in the two datasets is also shown in Figure 2.3. Since the two datasets are in different domains, the difference between the distribution of the facets is expected.

[10] http://tac.nist.gov/2014/BiomedSumm/
[11] National Institute of Standards and Technology
[12] https://github.com/WING-NUS/scisumm-corpus

Table 2.3: Characteristics of the datasets.

Characteristic	TAC	CL-SciSumm
# Documents	220	506
# Reference Documents	20	30
Avg. # Citing Docs for each Ref	15.5	15.9
Total # Citation Texts	313	702
Avg. Gold summary length (words)	235.6	134.2
Stdev. Gold summary length (words)	31.2	27.9
Separate train test sets	No	Yes

#: number of, Avg: average, and Stdev: standard deviation.

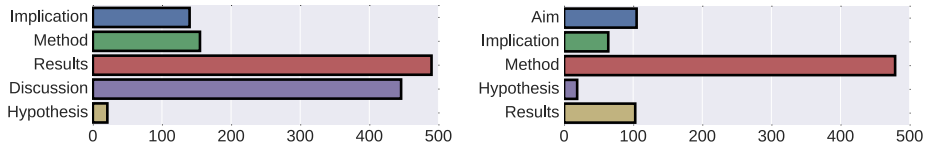


Figure 2.3: Distribution of discourse facets in each dataset.

2.2.4.2 Citation contextualization

Evaluation. Evaluation of the retrieved contexts is based on the overlap of the position of the retrieved contexts and the gold standard contexts. Per TAC guidelines[13], evaluation of the TAC benchmark was performed using character offset overlaps between the retrieved text spans and the annotated text spans. The overlap is weighted by the number of human annotators specifying gold spans. More formally, for a set of system retrieved contexts S, and gold standard context R =

[13]http://tac.nist.gov/2014/BiomedSumm/guidelines.html

$\{R_1 \cup R_2 \cup .... \cup R_m\}$ by m annotators, the weighted character based precision (P_{char}) and recall (R_{char}) are defined as follows:

$$P_{char} = \frac{\sum_i^m |S \cap R_i|}{m \times |S|} \qquad (2.9) \qquad\qquad R_{char} = \frac{\sum_i^m |S \cap R_i|}{\sum_i^m |R_i|} \qquad (2.10)$$

The official metric for the CL-SciSumm challenge was sentence level overlaps of the retrieved contexts with the gold standard. This was possible because unlike the articles in TAC which were in plaintext format, the sentence boundaries in CL-SciSumm were pre-specified. We also report character level metrics for the CL-SciSumm corpus; as we will see, the character level and sentence level metrics are more or less comparable.

One problem with position based evaluation metrics (character, or sentence) is that a system might retrieve a context that is in a different position than gold standard, but similar to the content of the gold standard. In such cases, the system is not rewarded at all. This is possible because authors might talk about a similar concept in different sections of the paper. To consider textual similarities of the retrieved context with the gold standard, we also compute ROUGE-N scores [155].

Comparison. To our knowledge, no review paper about the TAC challenge was released. Hence, for the TAC dataset, we compare our method against the following baselines which are standard well-known retrieval models suitable for this task:

- VSM. Ranking by Vector Space Model (VSM) with tf-idf weighting of the citations and the target reference contexts.
- BM25. BM25 scoring model [129] which is a probabilistic framework for ranking the relevant documents based on the query terms appearing in each document, regardless of their relative proximity.

- LMD. Language modeling with Dirichlet smoothing (LMD) [277] is a probabilistic framework that models the probability of documents generating the given query.

- LMD-LDA. An extension of the LMD retrieval model using Latent Dirichlet Allocation (LDA) which is recently proposed [125]. This model considers latent topics in ranking the relevant documents

For the CL-SciSumm data, we also compare against the top 5 best performing systems. For brief description about these approaches refer to section 2.2.2.

Results. The results on the TAC dataset are presented in Table 2.4. We observe that our proposed methods improve over all the baselines. Query Reformulation methods (NP and KW, respectively,) obtain character offset F1-scores of 23.8 and 24.1, which improve the best baseline by 7% and 8%. They also obtain higher ROUGE scores. This shows that noun phrases and key words can capture informative concepts in the citation that help better retrieving the related reference context. Our models based on word embeddings are also outperforming the baselines in virtually all metrics. General domain embeddings trained on Wikipedia (WE_{wiki}) and domain-specific embeddings trained on Genomics data (WE_{Bio}), achieve F1-scores of 23.2 and 25.5 with 4% and 14% improvement over the best baseline, respectively. Higher performance of the biomedical embeddings in comparison with general embeddings is expected because the words are captured in their correct context. An example is shown in Table 2.6, where the top similar words to the word "expression" are shown. The word "expression" in the biomedical context is defined as "the process by which genetic instructions are used to synthesize gene products". As we can see, using general domain embeddings, we might fail to capture this notion. Incorporating domain

Table 2.4: Results of citation contextualization on TAC 2014 dataset.

	Character offset overlap			Rouge	
Method	P_{char}	R_{char}	F_{char}	Rouge-2	Rouge-3
Baselines					
BM25 [225]	19.5	18.6	17.8	23.2	16.3
VSM	20.5	24.7	21.2	26.4	20.0
LMD [277]	21.3	26.7	22.3	27.2	20.8
LMD + LDA [125]	22.6	24.8	22.3	26.4	20.1
This work					
QR-Domain	24.1*	23.7	21.8	25.0	20.8
QR-NP	22.6	28.9*	23.8*	28.0*	21.8*
QR-KW	22.6	29.4*	24.1*	28.2*	22.2*
WE$_{wiki}$	21.8	28.5*	23.2*	26.9	20.9
WE$_{Bio}$	23.9*	31.2*	25.5*	29.2*	23.1*
WE$_{Bio+Retrofit}$	24.8*	**33.6***	26.4*	**30.7***	24.0*
WE$_{Bio}$ + *Domain*	**25.4***	33.0*	**27.0***	30.6*	**24.4***

The reported results are based on top 10 retrieved contexts. The top part shows the baselines and the bottom part shows our proposed model. Values are percentages. QR-Domain: Query Reformulation by Domain Ontology (UMLS), QR-NP: Query Reformulation by Noun Phrases, QR-KW: Query Reformulation by Key Words, WE$_{wiki}$: Word Embedding model with Wikipedia embeddings, WE$_{Bio}$: Word Embedding model with biomedical embeddings, WE$_{Bio+Retrofit}$: Incorporating domain knowledge in biomedical embeddings by retrofitting, WE$_{Bio}$ + *Domain*: Interpolated language model. * shows statistically significant improvement over all the baselines ($p<0.05$, t-test).

Table 2.5: Results of citation contextualization on CL-SciSumm 2016 dataset.

	Sentence overlap			ROUGE		Character offset overlap		
Method	P_{sent}	R_{sent}	F_{sent}	ROUGE-2	ROUGE-3	P_{char}	R_{char}	F_{char}
Other methods								
BM25 [225]	8.2	18.0	10.5	15.2	13.0	9.0	19.9	11.8
VSM	8.3	22.3	11.6	14.8	12.7	8.5	25.7	12.1
LM [277]	7.9	24.8	11.6	14.3	12.6	8.4	**26.1**	12.2
TSR [146]	5.3	4.7	5.0	-	-	-	-	-
Tf-idf + Neural Net [193]	9.2	11.1	10.0	-	-	-	-	-
SVM Rank [32]	8.8	13.1	10.3	-	-	-	-	-
Jaccard Fusion [153]	8.3	**26.1**	12.5	-	-	-	-	-
Tf-idf+stem [179]	9.6	22.4	13.4	-	-	-	-	-
This work								
QR-NP	8.8	20.4	12.2	15.8	13.6	9.7	23.8	13.2
QR-KW	9.0	21.3	12.6	16.0	13.8	9.6	23.3	13.0
WE_{wiki}	9.8	24.1	**13.9**	14.5	12.5	9.4	22.1	12.5
$WE_{wiki+Retrofit}$	9.8	23.8	13.8	14.7	13.6	8.2	22.3	12.0
Supervised	**11.3**	17.8	13.7	**17.5**	**15.0**	**12.0**	17.8	**13.7**

The reported values are percentages. The top part shows the baselines and state of the art models, while the bottom part shows our methods. P: Precision, R: Recall, F:F1-score. "sent" subscript shows overlap by sentences and "char" subscript shows character offset overlaps. QR-NP: Query Reformulation by Noun Phrases, QR-KW: Query Reformulation by Key Words, WE_{wiki}: Word Embedding model with Wikipedia embeddings, $WE_{wiki+Retrofit}$: Incorporating domain knowledge in embeddings by retrofitting. Results of our methods shown in bold are also significantly higher than that of all the three first baselines ($p<0.05$, t-test). Individual runs for other systems were not available to perform significance testing.

Table 2.6: The top similar words to a given sample word.

General (Wiki)	Domain-specific (Bio)
interpretation	upregulation
sense	mrna
emotion	protein
function	induction
show	cell

The words with highest similarity values to "expression" according to Word2Vec trained on Wikipedia (general domain) and Genomics collections (biomedical domain).

knowledge in the model results in further improvement as shown in last two rows of Table 2.4. The model using retrofitting $\text{WE}_{Bio+Retrofit}$ improves the best baseline by 18% while the interpolated model (WE_{Bio}+Domain) achieves the highest improvement by 21%. These results show the effectiveness of domain knowledge in the model.

Table 2.5 shows the results for the CL-SciSumm dataset. The first 3 rows are baselines that also are reported in TAC evaluation; in addition to those baselines, we also consider top performing state-of-the-art systems of 2016 CL-SciSumm (lines 4-8) as additional baselines to compare with. For the CL-SciSumm participating systems, we report the official sentence based evaluation metrics; the ROUGE scores and character based metrics were not reported in the official evaluation of the task. Some of our methods are specific to the biomedical domain such as WE_{Bio}; therefore, we do not evaluate those on the CL-SciSumm dataset which is in a completely different domain.

As shown in Table 2.5, our methods outperform the state-of-the-art on this dataset as well. The embedding-based model with Wikipedia trained embeddings (WE_{wiki})

achieves the best results with 13.9% F-1 score of sentence overlaps which is slightly higher than the F-1 score of 13.4 achieved by the best previous work (Tf-idf+stem in the Table) [179]. Interestingly, we observe that retrofitting ($WE_{wiki+Retrofit}$) does not improve over the standard embedding-based approach. This is likely due to the choice of the WordNet lexicon for retrofitting. While WordNet contains general domain terms, it does not necessarily capture relationships of words in the context of computational linguistics. In contrast to TAC where we had a domain specific lexicon suitable for the dataset, for the CL-SciSumm data we did not find any lexicon capturing the term relationships in the computational linguistics domain. We believe that retrofitting with such lexicon could result in further improvements. While query reformulation-based approaches improve over most of the baselines, their performance falls below the best baseline system. On the other hand, our supervised method also improves over the best baseline, achieving the highest overall prevision (11.3%) and ROUGE-2 (17.5%) and ROUGE-3 scores (15.0%).[14] It is encouraging that our embedding-based models (method names starting with "WE" in the Table 2.5), which are unsupervised models achieve the best results on this task and surpass the performance of the feature-rich supervised models in terms of sentence overlap. Table 2.7 shows the importance of each feature for our supervised method (explained in § 2.2.3.1). While the most important features are n-gram and character n-gram based tf-idf similarity, embedding based alignment and distance of average embeddings are also important in finding the correct context.

As evident from tables 2.4 and 2.5, the absolute system performances are not high, which further shows that this task is challenging. Since the TAC data are annotated by 4 people, we investigate the difficulty of this task for the human annotators. To do

[14]We do not report results of supervised model on TAC dataset because the TAC data do not have separate train and test sets.

Table 2.7: The weights (normalized) corresponding to the top features in the supervised method for citation contextualization (CL-SciSumm dataset).

Feature	weight
character n-gram tf-idf similarity	0.271
tf-idf similarity	0.201
embedding based alignment	0.189
distance average embeddings	0.106
bm25 similarity score	0.066
character n-gram count similarity	0.035
fuzzy word match	0.024
count based similarity	0.015
word match	0.013

Tf-idf similarity based features and embedding based features are the most helpful while the count based similarity and word matching features are among the least helpful features.

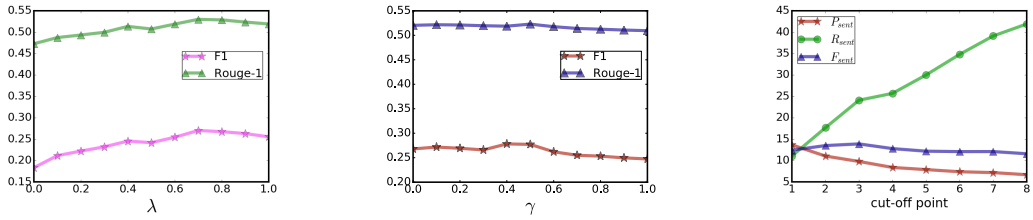


Figure 2.4: Parameters of the model for contextualization.

Table 2.8: Annotator agreement statistics.

Number of Citations	Number of Annotators with at least partial agreement
68	4
66	3
121	2
11	No agreement

The number of citations grouped by the number of annotators that agree at least partially on the context.

so, we calculate the agreement of the annotators with respect to the relevant context for the citations. Table 2.8 shows the number of citations grouped by the number of annotators that agree at least partially on the correct context. As illustrated, there are 68 citations out of 313 that all 4 annotators have partial agreement on the context span. This shows that the contextualization task is not trivial even for the human expert annotators.

Parameters. Our interpolated model of embeddings and domain knowledge (WE_{Bio}+Domain) has two main parameters γ and λ. Figure 2.4 shows the sensitivity of our model to different parameters. We observe that the best performance is achieved when $\gamma = 0.8$ and $\lambda = 0.5$. Our models retrieve a ranked list of contexts for the citations; we choose a cut-off point for returning the final results. Figure 2.4 also shows the effect of the cut-off point on one of our models.[15] We observe that the optimal cut-off point for best sentence F1-score is 3.

[15]The cut-off point has similar effect on all the models.

Table 2.9: Results for identifying the discourse facets for the retrieved contexts.

Method	P	R	F
Other methods			
SMO [232]	35.6	3.6	6.5
Decision tree [32]	59.7	9.0	15.3
Fusion method [153]	52.8	22.4	29.6
Jaccard cascade [153]	58.2	17.1	25.5
Jaccard Focused Method [153]	57.8	22.8	31.1
This work			
QR-NP	76.3	19.1	29.7
QR-KW	78.7	21.9	33.3
WE_{wiki}	82.7	22.4	33.1
$WE_{wiki+retro}$	81.7	23.4	34.8
Supervised	**83.1**	**23.7**	**36.1**

The metrics are Precision (P), Recall (R), and F1-score (F) of the identified discourse facets contingent on the correct retrieved span.

2.2.4.3 Identifying discourse facets

Evaluation. The official metric for evaluation of discourse facet identification is the Precision, Recall and F1-scores of the discourse facets, conditioned on the correctness of the retrieved reference context [121]. Therefore, we report the results for the CL-SciSumm data based on this metric. For the TAC dataset, the official metric is the classification accuracy weighted by the annotator agreements.[16] The accuracy for a system returned discourse facet is the number of annotators agreeing with that discourse facet divided by total number of annotators.

[16]http://tac.nist.gov/2014/BiomedSumm/guidelines/

Table 2.10: The classifier's intrinsic performance for identifying the discourse facets on the CL-SciSumm dataset.

Discourse Facet	P	R	F	#
Aim	0.93	0.36	0.52	36
Hypothesis	1.00	0.20	0.33	10
Implication	0.85	0.26	0.39	43
Method	0.79	0.98	0.87	250
Results	0.85	0.38	0.52	45
Average/Total	0.82	0.75	0.73	384

Results. Table 2.9 shows the results of our methods compared with the top performing official submitted runs to the CL-SciSumm 2016. We do not report the results of low performing systems. The classification algorithm for identifying the discourse facets is the method described in Section 2.2.3.2 across all our methods. However, since only the correct retrieved contexts are rewarded, the performance of each model differs based on the accuracy of retrieving the correct contexts. We observe that most of our methods (except for the QR-NP) improve over all the baselines in terms of all metrics. We obtain substantial improvements especially in terms of precision. The best method for identifying the discourse facets is the supervised method (indicated with "supervised" in the Table) which obtains 36.1% F-1 score, improving the best baseline ("Jaccard Focused Method") by 16%. Embedding methods also perform well by obtaining F-1 scores of 33.1% for the Wikipedia embeddings, and 34.8% for the retrofitted embeddings. These results further show the effectiveness of our contextualization methods along with the proposed classifier for identifying the facets.

Table 2.11: Effect of learning algorithms in identifying the discourse facets.

	SVM	RF	LR	Oracle
TAC	0.53	0.49	0.51	0.67
CL-SciSumm	0.67	0.64	0.66	-

SVM: Support Vector Machine with Linear Kernel, RF: Random Forest, LR: Logistic Regression, Oracle: Highest achievable score. Numbers are weighted accuracy scores by annotators.

We also demonstrate the intrinsic performance of our classifier for identifying the discourse facets in Table 2.10. As illustrated, the weighed average F1 performance over all discourse facets is 0.73. One challenge in identifying the discourse facets is the unbalanced dataset and the limited number of training examples for some specific facets. As also reflected in the table, we observe that for categories with smaller number of instances, the performance is generally lower. We therefore believe that having more training samples in the rare categories could further increase the performance.

Table 2.11 shows the results of facet identification in the TAC dataset as well as the effect of learning algorithms. Since for the TAC dataset there are 4 annotators, and the official metric is weighted accuracy scores, we also calculate the oracle score by always predicting what the majority of the annotators agree on. The oracle achieves 0.67 percent, suggesting that identifying discourse facets is not trivial for humans. We can see that the SVM classifier achieves the highest results with 81% relative accuracy to the oracle. For the CL-SciSumm dataset, there is only one annotator

Table 2.12: Summarization results on the CL-SciSumm dataset.

	ROUGE-2	ROUGE-3	ROUGE-SU4
LexRank [92]	11.8	8.1	11.4
CLexRank [216]	5.7	3.3	8.9
SumBasic [262]	8.5	3.8	11.5
SUMMA [232]	13.4	-	9.2
LMKL [61]	19.0	-	11.1
LMeq [61]	18.9	-	12.4
CIST [153]	21.9	-	13.6
QR-KW-iter	27.6	21.4	23.4
QR-KW-greedy	28.9	22.5	24.9
QR-NP-iter	23.0	20.9	22.6
QR-NP-greedy	**30.2**	**23.9**	**25.7**
WE_{wiki}-iter	22.4	15.9	21.7
WE_{wiki}-greedy	23.6	18.0	20.1
supervised-iter	24.1	18.5	20.8
supervised-greedy	23.6	18.3	19.6

Metrics are ROUGE F-scores. The top part shows the baselines and the state-of-the-art systems. Bottom systems show our method variants based on different contextualization approaches and sentence selection strategy from the discourse facets. *iter* (iterative) and *greedy* refer to the sentence selection approach for the final summary.

per discourse facet and therefore, the weighted accuracy metrics translates to simple accuracy scores.

To better analyze the effect of identifying discourse facets on the overall quality of the summary, we compare the ROUGE scores of the summary generated by our approach with and without this step. Table 2.14 shows the overall summarization results based on our QR-NP approach when we only use contextualized citations com-

pared with when we use faceted contextualized citations. We observe that grouping citation contexts by their corresponding discourse facet has a positive effect on the quality of the summary on both datasets (17% and 55% improvements over TAC and CL-SciSumm datasets in terms of ROUGE-2, respectively). This is because identifying facets and grouping the contextualized citations by facets, results in a summary that captures the content from all sections of the paper. We observe similar trends for other variants of our approaches; for brevity we only show the results for QR-NP as an illustrative analysis on the effect of identifying discourse facets on the quality of the generated summary.

Finally, an example of the generated summaries by our system (QR-NP-greedy) that uses citation contexts and discourse facets is illustrated in Figure 2.5. We observe that compared with the human summary, the summary generated by our system can capture the significant points of the paper.

2.2.4.4 SUMMARIZATION

We evaluate our summarization approach against the gold standard summaries written by human annotators. We set the summary length threshold to the average length of summary by words in each dataset (see Table 2.3). Table 2.12 shows the results for the summarization task. The first lines show the baselines which are existing summarization approaches including the SumBasic [262] algorithm and the original citation-based summarization approach [216]. The next four lines are the top state-of-the-art systems on the CL-SciSumm dataset. For the CL-SciSumm systems, the official reported results only included ROUGE-2 and ROUGE-SU4 scores. As illustrated in the table, virtually all our methods improve over the state-of-the-art, showing the effectiveness of our proposed summarization approach. Our best method (QR-NP-greedy) is based on the noun phrases query reformulation using the greedy

	Example summary
Human Summary	The limited coverage of lexical-semantic resources is a significant problem for NLP systems which can be alleviated by automatically classifying the unknown words. Supersense tagging assigns unknown nouns one of 26 broad semantic categories used by lexicographers to organise their manual insertion into WORDNET. Lexical-semantic resources have been applied successful to a wide range of Natural Language Processing (NLP) problems ranging from collocation extraction and class-based smoothing, to text classification and question answering. Some specialist topics are better covered in WORDNET than others. A considerable amount of research addresses structurally and statistically manipulating the hierarchy of WORDNET and the construction of new wordnet using the concept structure from English. Ciaramita and Johnson,implement a supersense tagger based on the multi-class preceptor classifier, which uses the standard collocation, spelling and syntactic features common in WSD and named entity recognition systems. The authors demonstrate the use of a very efficient shallow NLP pipeline to process a massive corpus. Such a corpus is needed to acquire reliable contextual information for the often very rare nouns they are attempting to supersense tag.
Our System	The limited coverage of lexical-semantic resources is a significant problem for NLP systems which can be alleviated by automatically classifying the unknown words. Ciaramita and Johnson (2003) present a tagger which uses synonym set glosses as annotated training examples. Our approach uses voting across the known supersenses of automatically extracted synonyms, to select a super- sense for the unknown nouns. The extracted synonyms are filtered before contributing to the vote with their supersense(s). Our development experiments are performed on the WORDNET 1.6 test set with one final run on the WORD- NET 1.7.1 test set. In particular, WORDNET (Fellbaum, 1998) has significantly influenced research in NLP. These results also support Ciaramita and JohnsonâĂŹs view that abstract concepts like communication, cognition and state are much harder. Lexicographers cannot possibly keep pace with language evolution: sense distinctions are continually made and merged, words are coined or become obsolete, and technical terms migrate into the vernacular.
CLexRank	Another related task is supersense tagging (Ciaramita and Johnson, 2003; Curran, 2005; Ciaramita and Altun, 2006). Supersense tagging (Ciaramita and Johnson, 2003; Curran, 2005) evaluates a model's ability to cluster words by their semantics. In contrast, some research have been focused on using predefined sets of sense-groupings for learning class-based classifiers for WSD. Although we could adapt our method for use with an automatically induced inventory, our method which uses WordNet might also be combined with one that can automatically find new senses from text and then relate these to WordNet synsets, as Ciaramita and Johnson (2003) and Curran (2005) do with unknown nouns.. An additional potential is to integrate automatically acquired relationships with the information found in WordNet, which seems to suffer from several serious limitations (Curran 2005), and typically overlaps to a rather limited extent with the output of automatic acquisition methods. Previous work on prediction at the supersense level (Ciaramita and Johnson, 2003; Curran, 2005) has focused on lexical acquisition (nouns exclusively), thus aiming at word type classification rather than tagging.

Figure 2.5: Example summaries. An example summary generated by our system (QR-NP-Greedy) on one of the papers from the CL-SciSumm dataset, compared with a human written summary and the output generated by CLexRank.

Table 2.13: Summarization results on the TAC dataset.

	ROUGE-2	ROUGE-3	ROUGE-SU4
LexRank [92]	12.8	5.0	17.5
CLexRank [216]	8.9	3.9	8.3
SumBasic [262]	8.3	4.2	12.5
QR-NP	**15.8**	**6.9**	20.4
QR-Domain	13.2	5.2	18.1
QR-KW	15.0	6.6	19.8
WE_{wiki}	13.3	5.5	17.8
WE_{Bio}	13.1	4.9	18.0
$WE_{Bio+Retrofit}$	14.4	5.7	19.5
WE_{Bio}+Domain	13.4	5.9	**20.7**

Metrics are ROUGE F-scores. The top part shows the baselines and the state-of-the-art systems. Bottom systems show our method variants based on different contextualization approaches and the greedy sentence selection strategy.

strategy of sentence selection. It achieves ROUGE-2 score of 30.2, which improves over the best baseline by 37.4%. In general, we can see that the greedy sentence selection strategy works better than the iterative approach. This is because the greedy strategy takes into account both the informativeness and the redundancy of the selected sentences.

Table 2.13 shows the results of summarization using on the TAC dataset. The reported approaches all use the greedy sentence selection strategy as it consistently outperforms the iterative approach. In general, while all our approaches outperform the baseline, query reformulation based approaches achieve the highest ROUGE scores; query reformulation method using noun phrases (QR-NP) achieves 15.8 and 6.9 ROUGE-2 and ROUGE-3 scores, respectively which is the highest scores. The interpo-

Table 2.14: The effect of discourse facets on the summarization results.

	R-2	R-3	R-SU4
TAC – QR-NP (no facet)	13.5	5.3	19.3
TAC – QR-NP (faceted)	15.8	6.9	20.4
CL-SciSumm – QR-NP (no facet)	19.4	17.2	22.6
CL-SciSumm – QR-NP (faceted)	30.2	23.9	25.7

The table shows the effect of discourse facets on the summarization results on the TAC and CL-SciSumm dataset based on QR-NP approach by greedy sentence selection strategy on the identified facets. Other approaches show similar positive trends. Metrics are ROUGE F-scores.

lated word embedding based model (WE_{Bio}+Domain) achieves the highest ROUGE-SU4 score (20.7). Comparing Tables 2.12 and 2.13 we notice that the scores for the TAC dataset are lower than that of CL-SciSum. This is due to the length of the generated summaries. As shown in Table 2.3, the average human summary length in the TAC data is almost 100 words more than the CL-SciSumm summaries. An interesting observation in these two tables is regarding the relative poor performance of the citation-based summarization baseline (CLexRank) that only uses citation texts in comparison with our methods that also take advantage of the citation context and the discourse structure of the articles. This observation further confirms our initial hypothesis that relying only on the citation texts could result in summaries that do not accurately reflect the content of the original paper, and that adding citation contexts can help produce better summaries.

2.2.5 Discussion

Citations are a significant part of scientific papers and analysis of citation texts can provide valuable information for various scholary applications. Our work provides new approaches for contextualizing citations which is a sub-task for enriching citation texts and thus can benefit various bibliometric enhanced NLP applications such as information extraction, information retrieval, article recommendation, and article summarization. Our work provides a comprehensive new framework for summarizing scientific papers that helps generating better scientific summaries.

We note that our evaluation was based on the Rouge automatic summarization evaluation framework. Automatic evaluation metrics have their own limitations and cannot fully characterize the effectiveness of the systems. Manual or semi-manual evaluation of summarization (e.g. through Pyramid framework) are alternative evaluation approaches that can provide additional insights into the performance of the systems. Yet, due to expense and reproduction issues, most of the standard evaluation benchmarks including TAC and CL-SciSumm have been evaluated through Rouge. As it is standard in the field and to be able to compare our results with the related work, we used the Rouge framework for evaluation. We also note that our focus has been on the content quality of the summaries and other criteria such as coherence and linguistic cohesion have not been the focus of our approach. Future work can investigate approaches for improving coherence and linguistic properties of the generated summaries.

2.3 A Discourse-Aware Attention Model for Abstractive Summarization of Scientific Documents

2.3.1 Introduction

There are two prominent approaches for document summarization. *(i)* Extractive approaches where the summary is generated by copying parts from the input; and *(ii)* abstractive approaches where the generated summary conveys the main aspects of the input document and it might include words or phrases that are not in the input document. The abstractive approach towards summarization is more similar to how human summarize documents [126]. In previous section, we presented an extractive approach for summarizing scientific papers which utilizes citations for capturing key contributions of a given document. Many scientific papers, however, do not contain sufficient number of citations. Similarly, newly published papers lack referencing papers. Motivated by these problems, we investigate models for directly summarizing a given scientific document without access to external information such as citations. Recently, sequence-to-sequence (seq2seq) neural network models [185, 207, 236] have achieved promising results in abstractive summarization. In these models, the document is fed to an encoder network and another (recurrent) network learns to decode the summary. However, these models typically have been used for summarizing short documents. For example, articles in the CNN/Daily Mail dataset [110] used in these works are on average about 600 words long. In contrast, scientific papers are much longer. Even short scientific papers include about 4 pages of content, while full conference papers and journal articles are significantly longer. Seq2seq models tend to struggle with longer sequences because at each decoding step, the decoder needs to learn to construct a context vector capturing relevant information from all the tokens in the source sequence [237]. One other distinct feature of scientific papers in com-

parison with existing summarization data is that scientific articles follow a standard discourse structure describing the problem, methodology, experiments/results, and finally conclusions [251].

Neural network models typically include a large number of parameters and training them requires large-scale datasets. Researchers have used large-scale news corpora such as CNN, Daily Mail and NY Times as summarization datasets where the articles are accompanied by a short abstract used as the ground truth summary.

In this section, we present an abstractive model for summarizing scientific papers which are an example of long-form structured document types. Our model includes a hierarchical encoder, capturing the discourse structure of the document and a discourse-aware decoder that generates the summary. Our decoder attends to different discourse sections and allows the model to better represent important information from the source resulting in a better context vector. We also introduce two large-scale datasets of long and structured scientific papers obtained from arXiv and PubMed to support both training and evaluating models on the task of long document summarization. Evaluation results show that our method outperforms state-of-the-art summarization models.

2.3.2 Background

In the seq2seq framework for abstractive summarization, an input document $\mathbf{x}$ is encoded using a Recurrent Neural Network (RNN) with $\mathbf{h}_i^{(e)}$ being the hidden state of the encoder at timestep i. The last step of the encoder is fed as input to another RNN which decodes the output one token at a time. Given an input document along with the corresponding ground-truth summary $\mathbf{y}$, the model is trained to output a summary $\hat{\mathbf{y}}$ that is close to $\mathbf{y}$. The output at timestep t is predicted using the decoder input $\mathbf{x}'_t$, decoder hidden state $\mathbf{h}_{t-1}^{(d)}$, and some information about the input sequence.

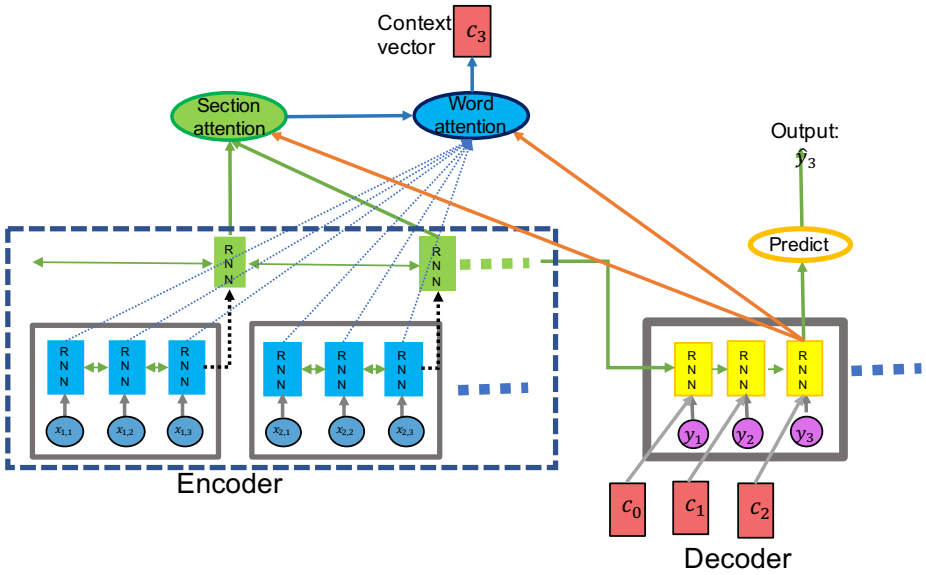


Figure 2.6: Overview of our model. The encoder has a hierarchical structure. The word-level RNN (blue) encodes discourse sections while another RNN (green) encodes the document. The decoder also consists of an RNN (yellow) and a "predict" network for generating the summary. At each decoding time step t ($t = 3$ is shown in the figure), the decoder forms a context vector c_t which is generated by attending to both sections and words. First the section attention weights (βs) are computed using the green "section attention" block. Then the word attention weights are computed using the blue "word attention" block. The context vector is used as another input to the decoder RNN and as an input to the "predict" network. The "predict" network, outputs the next word using a joint pointer-generator network.

This framework is the general seq2seq framework employed in many generation tasks including machine translation [11, 253] and summarization [46, 185].

2.3.2.1 Attentive decoding

The attention mechanism maps the decoder state and the encoder states to an output vector [263], which is a weighted sum of the encoder states and is called context vector

[11]. Incorporating this context vector at each decoding timestep (attentive decoding) is proven effective in seq2seq models. Formally, the context vector c_t is defined as:

$$\mathbf{c}_t = \sum_{i=1}^{N} \alpha_i^{(t)} \mathbf{h}_i^{(e)} \tag{2.11}$$

where $\alpha_i^{(t)}$ are weights calculated as follows:

$$\alpha_i^{(t)} = \text{softmax}(\text{score}(\mathbf{h}_i^{(e)}, \mathbf{h}_{t-1}^{(d)})) \tag{2.12}$$

The score function can be defined in bilinear, additive, and multiplicative ways [164]. We use the additive scoring function:

$$\text{score}(\mathbf{h}_i^{(e)}, \mathbf{h}_{t-1}^{(d)}) = \mathbf{v}_a^\top \tanh\left(\text{linear}(\mathbf{h}_i^{(e)}, \mathbf{h}_{t-1}^{(d)})\right) \tag{2.13}$$

where linear is a function that outputs a linear mapping of its arguments.

2.3.3 Model

We now describe our discourse-aware model (shown in Figure 2.6) for abstractive summarization of long documents.

2.3.3.1 Encoder

Our encoder extends the RNN encoder to a hierarchical RNN that captures the document discourse structure. We first encode each discourse section and then encode the document. Formally, we encode the document as a vector $\mathbf{d}$ according to the following:

$$\mathbf{d} = \text{RNN}_{doc}\left(\{\mathbf{s}_1, ..., \mathbf{s}_N\}\right) \tag{2.14}$$

where N is the number of sections in the document and $\mathbf{s}_i$ is the representation of section i in the document consisting of a sequence of tokens $\mathbf{x}_i$:

$$\mathbf{s}_i = \text{RNN}_{sec}\left(\{x_{(i,1)}, ...x_{(i,M)}\}\right) \tag{2.15}$$

where M is the maximum sequence length. RNN(.) denotes a function which is a recurrent neural network whose output is a vector representing the input sequence. The parameters of RNN_{sec} are shared for all the discourse sections. We use a single layer bidirectional LSTM (following the LSTM formulation of [104]) for both RNN_{doc} and RNN_{sec}; further extension to multilayer LSTM encoders is straightforward. We combine the forward and backward LSTM states by using a simple feed-forward network:

$$\mathbf{h} = \text{relu}(\mathbf{W}(\{\overrightarrow{\mathbf{h}}; \overleftarrow{\mathbf{h}}\} + \mathbf{b}) \tag{2.16}$$

where $\overrightarrow{\mathbf{h}}$, $\overleftarrow{\mathbf{h}}$, $\mathbf{W}$, and $\mathbf{b}$ respectively show the forward and backward LSTM networks, weights and biases.

2.3.3.2 Discourse-aware decoder

When humans summarize a long structured document, depending on the domain and the nature of the document, they try to capture the important points from the different discourse sections of the document. For example, scientific paper abstracts typically include the description of the problem, discussion of the methods, and finally results and conclusions [251]. Motivated by this observation, we propose a discourse-aware attention method. Intuitively, at each decoding timestep, in addition to the words in the document, we also attend to the relevant discourse section (filled orange circles in Figure 2.6). Then we use these discourse-related information to modify the

word-level attention function. Specifically, the context vector representing the source document is according to the following equation:

$$\mathbf{c}_t = \sum_{j=1}^{N} \sum_{i=1}^{M} \alpha_{(j,i)}^{(t)} \mathbf{h}_{(j,i)}^{(e)} \tag{2.17}$$

where $\mathbf{h}_{(j,i)}^{(e)}$ shows the encoder state of token i in discourse section j and $\alpha_{(j,i)}^{(t)}$ shows the corresponding weight to that encoder state. The weights $\alpha_{(j,i)}^{(t)}$ are obtained according to:

$$\alpha_{(j,i)}^{(t)} = \text{softmax}\left(\beta_j^{(t)} \text{score}(\mathbf{h}_{(j,i)}^{(e)}, \mathbf{h}_{t-1}^{(d)})\right) \tag{2.18}$$

The weights $\beta_j^{(t)}$ are updated according to:

$$\beta_j^{(t)} = \text{softmax}(\text{score}(\mathbf{s}_j, \mathbf{h}_{t-1}^{(d)})) \tag{2.19}$$

At each timestep t, the previous decoder state $\mathbf{h}_{t-1}^{(d)}$, and the context vector $\mathbf{c}_t$ are used to estimate the probability distribution of next word y_t:

$$p(y_t|y_{1:t-1}) = \text{softmax}\left(\mathbf{V}^\top \text{linear}\left(\mathbf{h}_{t-1}^{(d)}, \mathbf{c}_t\right)\right) \tag{2.20}$$

where $\mathbf{V}$ is a vocabulary weight matrix and linear is a linear mapping function. The input to the decoder RNN at each step t is the linear map between the context vector and the input; i.e., linear$(\mathbf{c}_t, \mathbf{x'}_t)$. Where $\mathbf{x'}_t$ is the gold standard token at training time and the previously predicted word at testing time.

2.3.3.3 Copying from source

There has been a surge of recent works in sequence learning tasks to address the problem of *unkown* token prediction by allowing the model to occasionally copy words

directly from source instead of generating a new token [106, 207, 236, 269]. Following these works, we add an additional binary variable z_t to the decoder, indicating generating a word from vocabulary (z_t=0) or copying a word from the source (z_t=1). The probability is learnt during training: $p(z_t{=}1|y_{1:t-1}) = \sigma(\text{linear}(\mathbf{h}_t^{(d)}, \mathbf{c}_t, \mathbf{x}'_t))$. Then the next word y_t is generated according to:

$$p(y_t|y_{1:t-1}) = \sum_z p(y_t, z_t{=}z|y_{1:t-1}); z = \{0, 1\} \tag{2.21}$$

The joint probability is decomposed as:

$$p(y_t, z_t{=}z) = \begin{cases} p_c(y_t|y_{1:t-1})\ p(z_t{=}z|y_{1:t-1}), & z{=}1 \\ p_g(y_t|y_{1:t-1})\ p(z_t{=}z|y_{1:t-1}), & z{=}0 \end{cases} \tag{2.22}$$

p_g is the probability of generating a word from the vocabulary and is defined according to Equation 2.20. p_c is the probability of copying a word from the source vector $\mathbf{x}$ and is defined as the sum of the word's attention weights. Specifically, the probability of copying a word x_ℓ is defined as:

$$p_c(y_t = x_\ell|y_{1:t-1}) = \sum\nolimits_{(j,i):x_{(j,i)}=x_\ell} \alpha^t_{(j,i)} \tag{2.23}$$

2.3.3.4 Decoder coverage

In long sequences, the neural generation models tend to repeat phrases where the softmax layer predicts the same phrase multiple times over multiple timesteps. In order to address this issue, following [236], we track attention coverage to prevent repeatedly attending to the same steps. This is done with a coverage vector $\mathbf{cov}^t$, the sum of attention weight vectors at previous timesteps: $\mathbf{cov}^t = \sum_{k=0}^{t-1} \alpha^k$. Note that this coverage also implicitly includes information about the attended document

Table 2.15: Statistics of our arXiv and PubMed datasets.

Datasets	# docs	avg doc. length (words)	avg. summary length (words)
CNN [185]	92K	656	43
Daily Mail [185]	219K	693	52
NY Times [207]	655K	530	38
PubMed (this work)	278K	6197	216
arXiv (this work)	194K	4938	220

discourse sections. We incorporate the decoder coverage as an additional input to the attention function:

$$\alpha_{(j,i)}^{(t)} = \text{softmax}\left(\beta_j^{(t)} \text{score}(\mathbf{h}_{(j,i)}^{(e)}, \text{cov}_{(j,i)}^{t}, \mathbf{h}_{t-1}^{(d)})\right)$$

2.3.4 Related work

Neural abstractive summarization models have been studied in the past [46, 185, 228] and later extended by source copying [170, 236], reinformcement learning [207], and sentence salience information [154]. One model variant of Nallapati et al. [185] is related to our model in using sentence-level information in attention, however, our model is different in encoding the document using a hierarchical encoder, using discourse sections in the decoding step, and utilizing a coverage mechanism. Similarly, in [157], the authors proposed a coarse-to-fine attention model that uses hard attention to find the text chunks of importance and then only attend to words in that chunk. In contrast, we consider all the discourse sections using soft attention. The closest model to ours is that of See et al. [236] and Paulus et al. [207] in using a joint pointer-generator network for summarization. However, our model extends theirs by

using *(i)* a hierarchical encoder for modeling long documents and *(ii)* a discourse-aware decoder that captures the information flow from all discourse sections of the document.

2.3.5 Experiments and results

2.3.5.1 Datasets

Seq2seq models typically have a large number of parameters and thus they require large training data with ground truth summaries. Researchers have constructed such training data from news artcicles (e.g. CNN, Daily Mail and New York Times articles), where the abstract of news articles is considered as ground truth summaries [185, 207]. However, news articles are relatively short and not suitable for the task of long-from document summarization. Following these works, we take scientific papers as an example of long documents with discourse information where their abstracts can be used as ground-truth summaries. We introduce two datasets collected from scientific repositories arXiv.org and PubMed.com.

The statistics of our datasets are shown in Table 2.15. In our datasets, both document and summary lengths are significantly larger than the existing large-scale summarization datasets. We retain 5% of this dataset as validation data, another 5% for test, and use the rest as the training set.

2.3.5.2 Dataset construction details

Scientific papers are examples of long documents that follow a standard discourse structure and they already come with ground truth summaries, making it possible to train supervised neural models. We follow existing work in constructing large-scale summarization datasets that take news article abstracts as ground truth.

We remove the documents that are excessively long (e.g. theses) or too short (e.g. tutorial announcement), not having an abstract or not having a discourse structure. We use the level-1 section headings as the discourse information. For arXiv, we use the LaTeX files and convert them to plain text using Pandoc[17] to preserve the discourse section information. We remove figures and tables using regular expressions to only preserve the textual information. We also normalize math formulas with a numbered special token *xmath-n* and replaced citation markers with *xcite*. We analyze the document section names and identify the most common concluding sections names (e.g. *conclusion*, *concluding remarks*, *summary*, etc). We only keep the sections up to the conclusion section of the document and we remove sections after the conclusion. This is done because we observe that sections succeeding the conclusion, are either acknowledgements, references, or supplemental/auxiliary material and do not usually convey any of the main points in the scientific paper.

2.3.5.3 Setup

Similar to the majority of published research in the summarization literature [46, 185, 236], evaluation was done using the Rouge automatic summarization evaluation metric [155] with full-length F-1 Rouge scores. We lowercase all tokens and perform sentence and word tokenization using spaCy [116].

2.3.5.4 Implementation details and model hyperparameters

We use tensorflow for implementing our models. We use the hyperparameters suggested by See et al. [236][18]. In particular, we use two bidirectional LSTMs with cell size of 256 and embedding dimensions of 128. Embeddings are trained from scratch

[17] https://pandoc.org/

[18] https://github.com/abisee/pointer-generator

and we did not find any gain using pre-trained embeddings. The vocabulary size is constrained to 50,000; using larger vocabulary size did not result in any improvement. We use mini-batches of size 16 and we limit the document length to 2500 and section length to 500 tokens. We use batch-padding and dynamic unrolling to handle variable sequence lengths in LSTMs. Training was done using Adagrad optimizer with learning rate 0.15 and an initial accumulator value of 0.1. The maximum decoder size was 210 tokens which is in line with average abstract length in our datasets. We first train the model without coverage and added it at the last two epochs to help the model converge faster. We train the models on NVIDIA Titan X Pascal GPUs. Training is performed for about 10 epochs and each training step takes about 3.2 seconds. We used beam search at decoding time with beam size of 4. We train the abstractive baselines for about 250K iterations as suggested by their authors.

2.3.5.5 Comparison

We compare our method with several well-known extractive baselines as well as state-of-the-art abstractive models using their open-sourced implementations, when available; we follow the exact training setup described in the corresponding papers. The compared methods are: *LexRank* [93], *SumBasic* [262], *LSA* [249], *Attn-Seq2Seq* [46, 185], *Pntr-Gen-Seq2Seq* [236]. The first three are extractive models and last two are abstractive. *Pntr-Gen-Seq2Seq* extends *Attn-Seq2Seq* by using a joint pointer network in decoding. For *Pntr-Gen-Seq2Seq* we use their reported hyperparameters so that the result differences are not due to hyperparameter tuning.

2.3.5.6 Results and discussion

Our main results are shown in Tables 2.16 and 2.17. Our method significantly outperforms both extractive and abstractive models, showing its effectiveness on both

Table 2.16: Results on the arXiv dataset.

Summarizer	RG-1	RG-2	RG-3	RG-L
Extractive				
SumBasic	29.47	6.95	2.36	26.30
LexRank	33.85	10.73	**4.54**	28.99
LSA	29.91	7.42	3.12	25.67
Abstractive				
Attn-Seq2Seq	29.30	6.00	1.77	25.56
Pntr-Gen-Seq2Seq	32.06	9.04	2.15	25.16
This work	**35.80***	**11.05**	3.62	**31.80***

RG stands for ROUGE. For our method, * shows statistically significant improvement over all the other methods ($p<0.05$, Wilcoxon signed-rank test).

datasets. We observe that in our ROUGE-1 score is respectively about 4 and 5 points higher than the abstractive model *Pntr-Gen-Seq2Seq* for the arXiv and PubMed datasets, providing a significant improvement. This shows that our extensions are effective for abstractive summarization of longer documents. The most competitive baseline method is *LexRank*, which is extractive. We note that since extractive methods copy salient sentences from the document it is usually easier for extractive methods to achieve better ROUGE scores in larger n-grams. Nevertheless, our method effectively outperforms *LexRank*. Figure 2.7 better shows how our model extensions are effective in capturing various discourse information from the papers. It can be observed that the state-of-the-art *Pntr-Gen-Seq2Seq* model generates a summary that mostly focuses on introducing the problem, however our model generates a summary that includes more information about the methodology and impacts of the target

Table 2.17: Results on the PubMed dataset.

Summarizer	RG-1	RG-2	RG-3	RG-L
Extractive				
SumBasic	34.91	12.81	8.13	31.74
LexRank	37.84	15.46	9.73	33.59
LSA	35.64	11.14	6.37	31.36
Abstractive				
Attn-Seq2Seq	30.70	9.45	7.19	25.56
Pntr-Gen-Seq2Seq	33.82	11.47	7.45	27.80
This work	**38.63***	**16.77***	**12.50***	**35.12***

RG stands for ROUGE. For our method, * shows statistically significant improvement over the other methods ($p<0.05$, Wilcoxon signed-rank test).

paper. This shows that the context vector in our model compared with *Pntr-Gen-Seq2Seq* is better able to capture important information from the source by attending to various discourse sections.

We used scientific papers as examples of long and structured documents and we showed that the model learns to effectively summarize such documents. While the majority of the scientific papers are associated with abstracts, there are indeed scientific writings whose abstract or summary are unavailable and we can use this system to summarize the key points of those articles. Examples include online scientific pages, blog posts, or some technical reports or white papers. Another interesting application of our model would be to generate scientific summaries that are longer than the abstracts. Our model is generic enough that given a dataset with ground-truth extended abstracts and discourse information, it can learn to generate those sum-

Abstract (Ground truth): in this paper , the author proposes a series of multilevel double hashing schemes called cascade hash tables . they use several levels of hash tables . in each table , we use the common double hashing scheme . higher level hash tables work as fail - safes of lower level hash tables . by this strategy , it could effectively reduce collisions in hash insertion . thus it gains a constant worst case lookup time with a relatively high load factor(@xmath0) in random experiments . different parameters of cascade hash tables are tested .

Pntr-Gen-Seq2Seq: hash table is a common data structure used in large set of data storage and retrieval . it has an o(1) lookup time on average , but the worst case lookup time can be as bad as . is the size of the hash table . we present a set of hash table schemes called cascade hash tables . hash table data structures which consist of several of hash tables with different size .

Our method: cascade hash tables are a common data structure used in large set of data storage and retrieval . such a time variation is essentially caused by possibly many collisions during keys hashing . in this paper , we present a set of hash schemes called cascade hash tables which consist of several levels(@xmath2) of hash tables with different size . after constant probes , if an item ca 'nt find a free slot in limited probes in any hash table , it will try to find a cell in the second level , or subsequent lower levels . with this simple strategy , these hash tables will have descendant load factors , therefore lower collision probabilities .

Figure 2.7: An example of the generated summary.

maries. Future work can design data collection methods to create large datasets with extended scientific abstracts. Such datasets could be used to train our model to generate extended abstracts.

We also note that following the convention in the summarization research, our quantitative evaluation is performed by ROUGE automatic metric. While ROUGE is an effective evaluation framework, nuances in the coherence or coverage of the summaries are not captured with it. It is non-trivial to evaluate such qualities especially for long

document summarization; future work can design expert human evaluations to explore these nuances.

2.4 Revisiting Summarization Evaluation for Scientific Articles

2.4.1 Introduction

Automatic text summarization has been an active research area in natural language processing for several decades. To compare and evaluate the performance of different summarization systems, the most intuitive approach is assessing the quality of the summaries by human evaluators. However, manual evaluation is expensive and the obtained results are subjective and difficult to reproduce [98]. To address these problems, automatic evaluation measures for summarization have been proposed. Rouge [155] is one of the first and most widely used metrics in summarization evaluation. It facilitates evaluation of system generated summaries by comparing them to a set of human written gold-standard summaries. It is inspired by the success of a similar metric Bleu [200] which is being used in Machine Translation (MT) evaluation. The main success of Rouge is due to its high correlation with human assessment scores on standard benchmarks [155]. Rouge has been used as one of the main evaluation metrics in later summarization benchmarks such as TAC[1] [198].

Since the establishment of Rouge, almost all research in text summarization have used this metric as the main means for evaluating the quality of the proposed approaches. The public availability of Rouge as a toolkit for summarization evaluation has contributed to its wide usage. While Rouge has originally shown good correlations with human assessments, the study of its effectiveness was only limited to a few benchmarks on news summarization data (DUC[2] 2001-2003 benchmarks). Since 2003, summarization has grown to much further domains and genres such as

[1]Text Analysis Conference (TAC) is a series of workshops for evaluating research in Natural Language Processing

scientific documents, social media and question answering. While there is not enough compelling evidence about the effectiveness of ROUGE on these other summarization tasks, published research is almost always evaluated by ROUGE. In addition, ROUGE has a large number of possible variants and the published research often (arbitrarily) reports only a few of these variants.

By definition, ROUGE solely relies on lexical overlaps (such as n-gram and sequence overlaps) between the system generated and human written gold-standard summaries. Higher lexical overlaps between the two show that the system generated summary is of higher quality. Therefore, in cases of terminology nuances and paraphrasing, ROUGE is not accurate in estimating the quality of the summary.

We study the effectiveness of ROUGE for evaluating scientific summarization. Scientific summarization targets much more technical and focused domains in which the goal is providing summaries for scientific articles. Scientific articles are much different than news articles in elements such as length, complexity and structure. Thus, effective summarization approaches usually have much higher compression rate, terminology variations and paraphrasing [257].

Scientific summarization has attracted more attention recently (examples include works by Abu-Jbara and Radev [2], Qazvinian et al. [219], and Cohan and Goharian [48]). Thus, it is important to study the validity of existing methodologies applied to the evaluation of news article summarization for this task. In particular, we raise the important question of how effective is ROUGE, as an evaluation metric for scientific summarization? We answer this question by comparing ROUGE scores with semi-manual evaluation score (Pyramid) in TAC 2014 scientific summarization dataset[1].

[2]Document Understanding Conference (DUC) was one of NIST workshops that provided infrastructure for evaluation of text summarization methodologies (`http://duc.nist.gov/`).

Results reveal that, contrary to the common belief, correlations between ROUGE and the Pyramid scores are weak, which challenges its effectiveness for scientific summarization. Furthermore, we show a large variance of correlations between different ROUGE variants and the manual evaluations which further makes the reliability of ROUGE for evaluating scientific summaries less clear. We then propose an evaluation metric based on relevance analysis of summaries which aims to overcome the limitation of high lexical dependence in ROUGE. We call our metric SERA (*Summarization Evaluation by Relevance Analysis*). Results show that the proposed metric achieves higher and more consistent correlations with semi-manual assessment scores.

Our contributions are as follows:

– Study the validity of ROUGE as the most widely-used summarization evaluation metric in the context of scientific summarization.

– Compare and contrast the performance of all variants of ROUGE in scientific summarization.

– Propose an alternative content relevance based evaluation metric for assessing the content quality of the summaries (SERA).

– Provide human Pyramid annotations for summaries in TAC 2014 scientific summarization dataset.[2]

2.4.2 Summarization evaluation by Rouge

ROUGE has been the most widely used family of metrics in summarization evaluation. In the following, we briefly describe the different variants of ROUGE:

[1] `http://www.nist.gov/tac/2014/BiomedSumm/`

[2] The annotations can be accessed via the following repository: `https://github.com/acohan/TAC-pyramid-Annotations/`

– ROUGE-N: ROUGE-N was originally a recall oriented metric that considered N-gram recall between a system generated summary and the corresponding gold human summaries. In later versions, in addition to the recall, precision was also considered in ROUGE-N, which is the precision of N-grams in the system generated summary with respect to the gold human summary. To combine both precision and recall, F1 scores are often reported. Common values of N range from 1 to 4.

– ROUGE-L: This variant of ROUGE compares the system generated summary and the human generated summary based on the Longest Common Subsequences (LCS) between them. The premise is that, longer LCS between the system and human summaries shows more similarity and therefore higher quality of the system summary.

– ROUGE-W: One problem with ROUGE-L is that all LCS with same lengths are rewarded equally. The LCS can be either related to a consecutive set of words or a long sequence with many gaps. While ROUGE-L treats all sequence matches equally, it makes sense that sequences with many gaps receive lower scores in comparison with consecutive matches. ROUGE-W considers an additional weighting function that awards consecutive matches more than non-consecutive ones.

– ROUGE-S: ROUGE-S computes the skip-bigram co-occurrence statistics between the two summaries. It is similar to ROUGE-2 except that it allows gaps between the bigrams by skipping middle tokens.

– ROUGE-SU: ROUGE-S does not give any credit to a system generated sentence if the sentence does not have any word pair co-occurring in the reference sentence. To solve this potential problem, ROUGE-SU was proposed which is an extension of ROUGE-S that also considers unigram matches between the two summaries.

ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU were later extended to consider both the recall and precision. In calculating ROUGE, stopword removal or stemming can also be considered, resulting in more variants.

In the summarization literature, despite the large number of variants of ROUGE, only one or very few of these variants are often chosen (arbitrarily) for evaluation of the quality of the summarization approaches. When ROUGE was proposed, the original variants were only recall-oriented and hence the reported correlation results [155]. The later extension of ROUGE family by precision were only reflected in the later versions of the ROUGE toolkit and additional evaluation of its effectiveness was not reported. Nevertheless, later published work in summarization adopted this toolkit for its ready implementation and relatively efficient performance.

The original ROUGE metrics show high correlations with human judgments of the quality of summaries on the DUC 2001-2003 benchmarks. However, these benchmarks consist of newswire data and are intrinsically very different than other summarization tasks such as summarization of scientific papers. We argue that ROUGE is not the best metric for all summarization tasks and we propose an alternative metric for evaluation of scientific summarization. The proposed alternative metric shows much higher and more consistent correlations with manual judgments in comparison with the well-established ROUGE.

2.4.3 SUMMARIZATION EVALUATION BY RELEVANCE ANALYSIS (SERA)

ROUGE functions based on the assumption that in order for a summary to be of high quality, it has to share many words or phrases with a human gold summary. However, different terminology may be used to refer to the same concepts and thus relying only on lexical overlaps may underrate content quality scores. To overcome this problem,

we propose an approach based on the premise that concepts take meanings from the context they are in, and that related concepts co-occur frequently.

Our proposed metric is based on analysis of the content relevance between a system generated summary and the corresponding human written gold-standard summaries. On high level, we indirectly evaluate the content relevance between the candidate summary and the human summary using information retrieval. To accomplish this, we use the summaries as search queries and compare the overlaps of the retrieved results. Larger number of overlaps, suggest that the candidate summary has higher content quality with respect to the gold-standard. This method, enables us to also reward for terms that are not lexically equivalent but semantically related. Our method is based on the well established linguistic premise that semantically related words occur in similar contexts [261]. The context of the words can be considered as surrounding words, sentences in which they appear or the documents. For scientific summarization, we consider the context of the words as the scientific articles in which they appear. Thus, if two concepts appear in identical set of articles, they are semantically related. We consider the two summaries as similar if they refer to same set of articles even if the two summaries do not have high lexical overlaps. To capture if a summary relates to a article, we use information retrieval by considering the summaries as queries and the articles as documents and we rank the articles based on their relatedness to a given summary. For a given pair of system summary and the gold summary, similar rankings of the retrieved articles suggest that the summaries are semantically related, and thus the system summary is of higher quality.

Based on the domain of interest, we first construct an index from a set of articles in the same domain. Since TAC 2014 was focused on summarization in the biomedical domain, our index also comprises of biomedical articles. Given a candidate summary C and a set of gold summaries G_i ($i = 1, ..., M$; M is the total number of human

summaries), we submit the candidate summary and gold summaries to the search engine as queries and compare their ranked results. Let $I = \langle d_1, ..., d_N \rangle$ be the entire index which comprises of N total documents.

Let $R_C = \langle d_{\ell_1}, ..., d_{\ell_n} \rangle$ be the ranked list of retrieved documents for candidate summary C, and $R_{G_i} = \langle d_{\ell_1^{(i)}}, ..., d_{\ell_n^{(i)}} \rangle$ the ranked list of results for the gold summary G_i. These lists of results are based on a rank cut-off point n that is a parameter of the system. We provide evaluation results on different choices of cut-off point n in the Section 2.4.5 We consider the following two scores: (i) simple intersection and (ii) discounted intersection by rankings. The simple intersection just considers the overlaps of the results in the two ranked lists and ignores the rankings. The discounted ranked scores, on the other hand, penalizes ranking differences between the two result sets. As an example consider the following list of retrieved documents (denoted by d_is) for a candidate and a gold summary as queries:

Results for candidate summary: $\langle d_1,\ d_2,\ d_3,\ d_4 \rangle$

Results for gold summary: $\langle d_3,\ d_2,\ d_1,\ d_4 \rangle$

These two sets of results consist of identical documents but the ranking of the retrieved documents differ. Therefore, the simple intersection method assigns a score of 1.0 while in the discounted ranked score, the score will be less than 1.0 (due to ranking differences between the result lists).

We now define the metrics more precisely. Using the above notations, without loss of generality, we assume that $|R_C| \geq |R_{G_i}|$. SERA is defined as follows:

$$\text{SERA} = \frac{1}{M} \sum_{i=1}^{M} \frac{|R_C \cap R_{G_i}|}{|R_C|}$$

To also account for the ranked position differences, we modify this score to discount rewards based on rank differences. That is, in ideal score, we want search results from candidate summary (R_C) to be the same as results for gold-standard summaries (R_G)

and the rankings of the results also be the same. If the rankings differ, we discount the reward by log of the differences of the ranks. More specifically, the discounted score (SERA-DIS) is defined as:

$$\text{SERA-DIS} = \frac{\sum_{i=1}^{M} \Big(\sum_{j=1}^{|R_C|} \sum_{k=1}^{|R_{G_i}|} \begin{cases} \left(\frac{1}{\log(|j-k|+2)}\right) & \text{if } R_C^{(j)} = R_{G_i}^{(k)} \\ 0 & \text{otherwise} \end{cases} \Big)}{M \times D_{\max}}$$

where, as previously defined, M, R_C and R_{G_i} are total number of human gold summaries, result list for the candidate summary and result list for the human gold summary, respectively. In addition, $R_C^{(j)}$ shows the jth results in the ranked list R_C and D_{max} is the maximum attainable score used as the normalizing factor.

We use elasticsearch[1], an open-source search engine, for indexing and querying the articles. For retrieval model, we use the Language Modeling retrieval model with Dirichlet smoothing [276]. Since TAC 2014 benchmark is on summarization of biomedical articles, the appropriate index would be the one constructed from articles in the same domain. Therefore, we use the open access subset of Pubmed[2] which consists of published articles in biomedical literature.

We also experiment with different query (re)formulation approaches. Query reformulation is a method in Information Retrieval that aims to refine the query for better retrieval of results. Query reformulation methods often consist of removing ineffective terms and expressions from the query (query reduction) or adding terms to the query that help the retrieval (query expansion). Query reduction is specially impor-

[1] `https://github.com/elastic/elasticsearch`

[2] PubMed is a comprehensive resource of articles and abstracts published in life sciences and biomedical literature `http://www.ncbi.nlm.nih.gov/pmc/`

tant when queries are verbose. Since we use the summaries as queries, the queries are usually long and therefore we consider query reductions.

In our experiments, the query reformulation is done by 3 different ways: (*i*) Plain: The entire summary without stopwords and numeric values; (*ii*) Noun Phrases (NP): We only keep the noun phrases as informative concepts in the summary and eliminate all other terms[19]; and (*iii*) Keywords (KW): We only keep the keywords and key phrases in the summary. For extracting the keywords and keyphrases (with length of up to 3 terms), we extract expressions whose *idf*[1] values is higher than a predefined threshold that is set as a parameter. We set this threshold to the average *idf* values of all terms except stopwords. *idf* values are calculated on the same index that is used for the retrieval.

We hypothesize that using only informative concepts in the summary prevents query drift and leads to retrieval of more relevant documents. Noun phrases and keywords are two heuristics for identifying the informative concepts.

2.4.4 Experiments

2.4.4.1 Data

To the best of our knowledge, the only scientific summarization benchmark is from TAC 2014 summarization track. For evaluating the effectiveness of Rouge variants and our metric (Sera), we use this benchmark, which consists of 20 topics each with a biomedical journal article and 4 gold human written summaries.

[19]We use `https://spacy.io/` parser to detect noun phrases

[1]Inverted Document Frequency

2.4.4.2 Annotations

In the TAC 2014 summarization track, ROUGE was suggested as the evaluation metric for summarization and no human assessment was provided for the topics. Therefore, to study the effectiveness of the evaluation metrics, we use the semi-manual Pyramid evaluation framework [187, 188]. In the pyramid scoring, the content units in the gold human written summaries are organized in a pyramid. In this pyramid, the content units are organized in tiers and higher tiers of the pyramid indicate higher importance. The content quality of a given candidate summary is evaluated with respect to this pyramid.

To analyze the quality of the evaluation metrics, following the pyramid framework, we design an annotation scheme that is based on identification of important content units. Consider the following example:

Endogeneous small RNAs (miRNA) were genetically screened and studied to find the miRNAs which are related to tumorigenesis.

In the above example, the underlined expressions are the content units that convey the main meaning of the text. We call these small units, nuggets which are phrases or concepts that are the main contributors to the content quality of the summary.

We asked two human annotators to review the gold summaries and extract content units in these summaries. The pyramid tiers represent the occurrences of nuggets across all the human written gold-standard summaries, and therefore the nuggets are weighted based on these tiers. The intuition is that, if a nugget occurs more frequently in the human summaries, it is a more important contributor (thus belongs to higher tier in the pyramid). Thus, if a candidate summary contains this nugget, it should be rewarded more. An example of the nuggets annotations in pyramid framework is shown in Table 2.18. In this example, the nugget "*cell mutation*" belongs to the 4th

Table 2.18: Example of nugget annotation for Pyramid scores.

id	nugget	Tier
n_1	IDH1/2	3
n_2	isocitrate dehydrogenase 1 & 2	2
n_3	alpha ketoglutarate-dependent enzyme	1
n_4	TET2	1
n_5	cell mutation	4
n_6	DNA methylation	2

The pyramid tier represents the number of occurrences of the nugget in all the human written gold summaries.

tier and it suggests that the "*cell mutation*" nugget is a very important representative of the content of the corresponding document.

Let T_i define the tiers of the pyramid with T_1 being the bottom tier and T_n the top tier. Let N_i be the number of the nuggets in the candidate summary that appear in the tier T_i. Then the pyramid score P of the candidate summary will be:

$$P = \frac{1}{P_{\max}} \sum_{i=1}^{n} i \times N_i$$

where $P_{\max}$ is the maximum attainable score used for normalizing the scores:

$$P_{\max} = \sum_{i=j+1}^{n} i \times |T_i| + j \times (X - \sum_{i=j+1}^{n} |T_i|)$$

where X is the total number of nuggets in the summary and $j = \max_i \sum_{t=i}^{n} |T_t| \geq X$.

We release the pyramid annotations of the TAC 2014 dataset through a public repository[2].

[2] `https://github.com/acohan/TAC-pyramid-Annotations`

2.4.4.3 Summarization approaches

We study the effectiveness of ROUGE and our proposed method (SERA) by analyzing the correlations with semi-manual human judgments. Very few teams participated in TAC 2014 summarization track and the official results and the review paper of TAC 2014 systems were never published. Therefore, to evaluate the effectiveness of ROUGE, we applied 9 well-known summarization approaches on the TAC 2014 scientific summarization dataset. Obtained ROUGE and SERA results of each of these approaches are then correlated with semi-manual human judgments. In the following, we briefly describe each of these summarization approaches.

1. LexRank [92]: LexRank finds the most important (central) sentences in a document by using random walks in a graph constructed from the document sentences. In this graph, the sentences are nodes and the similarity between the sentences determines the edges. Sentences are ranked according to their importance. Importance is measured in terms of centrality of the sentence — the total number of edges incident on the node (sentence) in the graph. The intuition behind LexRank is that a document can be summarized using the most central sentences in the document that capture its main aspects.

2. Latent Semantic Analysis (LSA) based summarization [248]: In this summarization method, Singular Value Decomposition (SVD) [82] is used for deriving latent semantic structure of the document. The document is divided into sentences and a term-sentence matrix $\mathbf{A}$ is constructed. The matrix $\mathbf{A}$ is then decomposed into a number of linearly-independent singular vectors which represent the latent concepts in the document. This method, intuitively, decomposes the document into several latent topics and then selects the most representative sentences for each of these topics as the summary of the document.

3. Maximal Marginal Relevance (MMR) [33]: Maximal Marginal Relevance (MMR) is a greedy strategy for selecting sentences for the summary. Sentences are added iteratively to the summary based on their relatedness to the document as well as their novelty with respect to the current summary.

4. Citation based summarization [219]: In this method, citations are used for summarizing an article. Using the LexRank algorithm on the citation network of the article, top sentences are selected for the final summary.

5. Using frequency of the words [163]: In this method, which is one the earliest works in text summarization, raw word frequencies are used to estimate the saliency of sentences in the document. The most salient sentences are chosen for the final summary.

6. SumBasic [262]: SumBasic is an approach that weights sentences based on the distribution of words that is derived from the document. Sentence selection is applied iteratively by selecting words with highest probability and then finding the highest scoring sentence that contains that word. The word weights are updated after each iteration to prevent selection of similar sentences.

7. Summarization using citation-context and discourse structure [48]: In this method, the set of citations to the article are used to find the article sentences that directly reflect those citations (citation-contexts). In addition, the scientific discourse of the article is utilized to capture different aspects of the article. The scientific discourse usually follows a structure in which the authors first describe their hypothesis, then the methods, experiment, results and implications. Sentence selection is based on finding the most important sentences in each of the discourse facets of the document using the MMR heuristic.

8. KL Divergence [108] In this method, the document unigram distribution P and the summary unigram distributation Q are considered; the goal is to find a summary whose distribution is very close to the document distribution. The difference of the distributions is captured by the Kullback-Lieber (KL) divergence, denoted by $\mathrm{KL}(P||Q)$.

9. Summarization based on Topic Models [108]: Instead of using unigram distributions for modeling the content distribution of the document and the summary, this method models the document content using an LDA based topic model [22]. It then uses the KL divergence between the document and the summary content models for selecting sentences for the summary.

2.4.5 Results and discussion

We calculated all variants of Rouge scores, our proposed metric, Sera, and the Pyramid score on the generated summaries from the summarizers described in Section 2.2.2.3. We do not report the Rouge, Sera or pyramid scores of individual systems as it is not the focus of this study. Our aim is to analyze the effectiveness of the evaluation metrics, not the summarization approaches. Therefore, we consider the correlations of the automatic evaluation metrics with the manual Pyramid scores to evaluate their effectiveness; the metrics that show higher correlations with manual judgments are more effective.

Table 2.19 shows the Pearson, Spearman and Kendall correlation of Rouge and Sera, with pyramid scores. Both Rouge and Sera are calculated with stopwords removed and with stemming. Our experiments with inclusion of stopwords and without stemming showed similar results and thus, we do not include those to avoid redundancy.

Table 2.19: Correlation between ROUGE and SERA.

	Pyramid		
Metric	Pearson(r)	Spearman(ρ)	Kendall(τ)
ROUGE-1-F	0.454	0.174	0.138
ROUGE-1-P	0.257	0.116	0
ROUGE-1-R	0.513	0.229	0.138
ROUGE-2-F	0.816	0.696	0.552
ROUGE-2-P	0.824	0.841	0.69
ROUGE-2-R	0.803	0.696	0.552
ROUGE-3-F	**0.878**	0.841	0.69
ROUGE-3-P	0.875	0.725	0.552
ROUGE-3-R	0.875	0.841	0.69
ROUGE-L-F	0.454	0.261	0.276
ROUGE-L-P	0.262	0.29	0.138
ROUGE-L-R	0.52	0.261	0.276
ROUGE-S-F	0.603	0.406	0.414
ROUGE-S-P	0.344	0.174	0.138
ROUGE-S-R	0.664	0.406	0.414
ROUGE-SU-F	0.601	0.493	0.462
ROUGE-SU-P	0.338	0.174	0.138
ROUGE-SU-R	0.662	0.406	0.414
ROUGE-W-1.2-F	0.607	0.493	0.414
ROUGE-W-1.2-P	0.418	0.377	0.276
ROUGE-W-1.2-R	0.626	0.667	0.552
SERA-5	0.823	0.941	0.857
SERA-10	0.788	0.647	0.429
SERA-KW-5	0.848	0.765	0.571
SERA-KW-10	0.641	0.618	0.486
SERA-NP-5	0.859	**1.0**	**1.0**
SERA-NP-10	0.806	0.941	0.857
SERA-DIS-5	0.631	0.824	0.714
SERA-DIS-10	0.687	0.824	0.714
SERA-DIS-KW-5	0.838	0.941	0.857
SERA-DIS-KW-10	0.766	0.712	0.729
SERA-DIS-NP-5	0.834	0.941	0.857
SERA-DIS-NP-10	0.86	0.941	0.857

All variants of ROUGE are displayed. F: F-Score; R: Recall; P: Precision; DIS: Discounted variant of SERA; KW: using Keyword query reformulation; NP: Using noun phrases for query reformulation. The numbers in front of the SERA metrics indicate the rank cut-off point.

2.4.5.1 SERA

The results of our proposed method (SERA) are shown in the bottom part of Table 2.19. In general, SERA shows better correlation with pyramid scores in comparison with ROUGE. We observe that the Pearson correlation of SERA with cut-off point of 5 (shown by SERA-5) is 0.823 which is higher than most of the ROUGE variants. Similarly, the Spearman and Kendall correlations of the SERA evaluation score is 0.941 and 0.857 respectively, which are higher than all ROUGE correlation values. This shows the effectiveness of the simple variant of our proposed summarization evaluation metric.

Table 2.19 also shows the results of other SERA variants including discounting and query reformulation methods. Some of these variants are the result of applying query reformulation in the process of document retrieval which are described in Section 2.4.3 As illustrated, the Noun Phrases (NP) query reformulation at cut-off point of 5 (shown as SERA-NP-5) achieves the highest correlations among all the SERA variants ($r = 0.859$, $\rho = \tau = 1.0$). In the case of Keywords (KW) query reformulation, without using discounting, we can see that there is no positive gain in correlation. However, keywords when applied on the discounted variant of SERA, result in higher correlations.

Discounting has more positive effect when applied on query reformulation-based SERA than on the simple variant of SERA. In the case of discounting and NP query reformulation (SERA-DIS-NP), we observe higher correlations in comparison with simple SERA. Similarly, in the case of Keywords (KW), positive correlation gain is obtained in most of correlation coefficients. NP without discounting and at cut-off point of 5 (SERA-NP-5) shows the highest non-parametric correlation. In addition, the

discounted NP at cut-off point of 10 (SERA-NP-DIS-10) shows the highest parametric correlations.

In general, using NP and KW as heuristics for finding the informative concepts in the summary effectively increases the correlations with the manual scores. Selecting informative terms from long queries results in more relevant documents and prevents query drift. Therefore, the overall similarity between the two summaries (candidate and the human written gold summary) is better captured.

2.4.5.2 ROUGE

Another important observation is regarding the effectiveness of ROUGE scores (top part of Table 2.19). Interestingly, we observe that many variants of ROUGE scores do not have high correlations with human pyramid scores. The lowest F-score correlations are for ROUGE-1 and ROUGE-L (with r=0.454). Weak correlation of ROUGE-1 shows that matching unigrams between the candidate summary and gold summaries is not accurate in quantifying the quality of the summary. On higher order n-grams, however, we can see that ROUGE correlates better with pyramid. In fact, the highest overall r is obtained by ROUGE-3. ROUGE-L and its weighted version ROUGE-W, both have weak correlations with pyramid. Skip-bigrams (ROUGE-S) and its combination with unigrams (ROUGE-SU) also show sub-optimal correlations. Note that ρ and τ correlations are more reliable in our setup due to the small sample size.

These results confirm our initial hypothesis that ROUGE is not accurate estimator of the quality of the summary in scientific summarization. We attribute this to the differences of scientific summarization with general domain summaries. When humans summarize a relatively long research paper, they might use different terminology and paraphrasing. Therefore, ROUGE which only relies on term matching between

a candidate and a gold summary, is not accurate in quantifying the quality of the candidate summary.

Table 2.20: Correlation between SERA and ROUGE scores.

	ROUGE-2-F			ROUGE-3-F		
Metric	r	ρ	τ	r	ρ	τ
SERA-5	.408	.522	.414	.540	.725	.552
SERA-10	.447	.406	.276	0.6	.667	.414
SERA-KW-5	.867	.754	.690	.770	.899	.828
SERA-KW-10	.574	.174	.138	.343	.029	0
SERA-NP-5	.588	.696	.552	.720	.841	.690
SERA-NP-10	.416	.522	.414	.609	.725	.552
SERA-DIS-5	.154	.464	.276	.396	.667	.414
SERA-DIS-10	.280	.464	.276	.502	.667	.414
SERA-DIS-KW-5	.891	.812	.690	.842	.899	.828
SERA-DIS-KW-10	.751	.696	.552	.650	.551	.414
SERA-DIS-NP-5	.584	.522	.414	.744	.725	.552
SERA-DIS-NP-10	.583	.522	.414	.763	.725	.552

NP: Query reformulation with Noun Phrases; KW: Query reformulation with Keywords; DIS: Discounted variant of SERA; The numbers in front of the SERA metrics indicate the rank cut-off point.

2.4.5.3 CORRELATION OF SERA WITH ROUGE

Table 2.20 shows correlations of our metric SERA with ROUGE-2 and ROUGE-3, which are the highest correlated ROUGE variants with pyramid. We can see that in general, the correlation is not strong. Keyword based reduction variants are the only variants for which the correlation with ROUGE is high. Looking at the correlations of KW variants of SERA with pyramid (Table 2.19, bottom part), we observe that these variants are also highly correlated with manual evaluation.

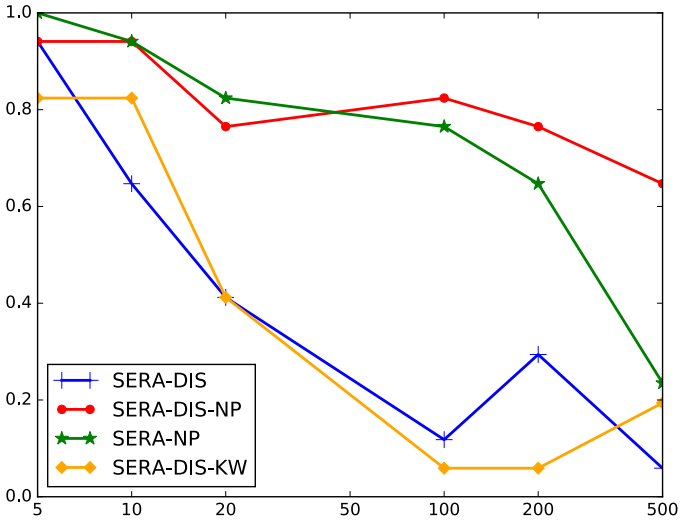


Figure 2.8: ρ correlation of SERA with pyramid based on different cut-off points. The x-axis shows the cut-off point parameter. DIS: Discounted variant of SERA; NP: Query reformulation with Noun Phrases; KW: Query reformulation with Keywords.

2.4.5.4 EFFECT OF THE RANK CUT-OFF POINT

Finally, Figure 2.8 shows ρ correlation of different variants of SERA with pyramid based on selection of different cut-off points (r and τ correlations result in very similar graphs). When the cut-off point increases, more documents are retrieved for the candidate and the gold summaries, and therefore the final SERA score is more fine-grained. A general observation is that as the search cut-off point increases, the correlation with pyramid scores decreases. This is because when the retrieved result list becomes larger, the probability of including less related documents increases which negatively affects correct estimation of the similarity of the candidate and gold summaries. The

most accurate estimations are for metrics with cut-off points of 5 and 10 which are included in the reported results of all variants in Table 2.19.

2.4.6 Related work

ROUGE [155] assesses the content quality of a candidate summary with respect to a set of human gold summaries based on their lexical overlaps. ROUGE consists of several variants. Since its introduction, ROUGE has been one of the most widely reported metrics in the summarization literature, and its high adoption has been due to its high correlation with human assessment scores in DUC datasets [155]. However, later research has casted doubts about the accuracy of ROUGE against manual evaluations. In [60], the authors analyzed DUC 2005 to 2007 data and showed that while some systems achieve high ROUGE scores with respect to human summaries, the linguistic and responsiveness scores of those systems do not correspond to the high ROUGE scores.

We studied the effectiveness of ROUGE through correlation analysis with manual scores. Besides correlation with human assessment scores, other approaches have been explored for analyzing the effectiveness of summarization evaluation. Rankel et al. [220] studied the extent to which a metric can distinguish between the human and system generated summaries. They also proposed the use of paired two-sample t-tests and the Wilcoxon signed-rank test as an alternative to ROUGE in evaluating several summarizers. Similarly, Owczarzak et al. [199] proposed the use of multiple binary significance tests between the system summaries for ranking the best summarizers.

Since introduction of ROUGE, there have been other efforts for improving automatic summarization evaluation. Hovy et al. [118] proposed an approach based on comparison of so called Basic Elements (BE) between the candidate and reference summaries. BEs were extracted based on syntactic structure of the sentence. The

work by Conroy et al. [63] was another attempt for improving ROUGE for update summarization which combined two different ROUGE variants and showed higher correlations with manual judgments for TAC 2008 update summaries.

Apart from the content, other aspects of summarization such as linguistic quality have been also studied. Pitler et al. [212] evaluated a set of models based on syntactic features, language models and entity coherences for assessing the linguistic quality of the summaries. Machine translation evaluation metrics such as BLEU have also been compared and contrasted against ROUGE [101]. Despite these works, when gold-standard summaries are available, ROUGE is still the most common evaluation metric that is used in the summarization published research. Apart from ROUGE's initial good results on the newswire data, the availability of the software and its efficient performance have further contributed to its popularity.

2.4.7 DISCUSSION

Our analysis on the effectiveness of evaluation measures for scientific summaries was performed using correlations with manual judgments. An alternative approach to follow would be to use statistical significance testing on the ability of the metrics to distinguish between the summarizers (similar to Rankel et al. [220]). We studied the effectiveness of existing summarization evaluation metrics in the scientific text genre and proposed an alternative superior metric. Another extension of this work would be to evaluate automatic summarization evaluation in other genres of text (such as social media). Our proposed method only evaluates the content quality of the summary. Similar to most of existing summarization evaluation metrics, other qualities such as linguistic cohesion, coherence and readability are not captured by this method. Developing metrics that also incorporate these qualities is yet another future direction to follow.

2.5 Conclusions

I first presented an extractive unified framework for scientific document summarization; the framework consisted of three main parts: finding the context for the citations in the reference paper, identifying the discourse facet of each citation context, and generating the summary from the faceted citation contexts. I utilized query reformulation methods, word embeddings, and domain knowledge in our methods to capture the terminology variations between the citing and cited authors. I demonstrated the effectiveness of this approach on two scientific document summarization benchmarks each in a different domain. I improved over the state-of-the-art by large margins in most of the tasks. While the results are encouraging, the absolute values of some metrics especially in the contextualization task suggest that this problem is worth further exploration. Contextualizing citations is a new task and not only it helps improving scientific document summarization, but also it can benefit other bibliometric enhanced end-to-end applications such as keyword extraction, information retrieval, and article recommendation.

I also presented an abstractive summarization method that could be viewed as a complementary approach to my extractive method especially for the cases where the paper does not have enough citations. Most existing successful approaches in summarizing long documents are extractive in nature where the summary is formed by copying important parts of the input document. Instead, I approached the scientific document summarization problem in an abstractive fashion. My goal was to generate a summary that is written from scratch and not necessarily including the exact sentences or phrases in the article itself. I presented a neural discourse-aware sequence-to-sequence model that is able to effectively summarize long and structured documents such as scientific papers. I showed how our methods significantly improve

over the state-of-the-art abstractive methods in terms of ROUGE evaluation metrics. I also introduced two new large-scale datasets of scientific papers for supporting training and evaluating systems on the task of long document summarization. These datasets can help the community to further explore this problem.

Finally, this chapter provided an analysis of existing evaluation metrics for scientific summarization with evaluation of all variants of ROUGE. I showed that ROUGE may not be the best metric for summarization evaluation; especially in summaries with high terminology variations and paraphrasing (e.g. scientific summaries). Furthermore, I showed that different variants of ROUGE result in different correlation values with human judgments, indicating that not all ROUGE scores are equally effective. Among all variants of ROUGE, ROUGE-2 and ROUGE-3 better correlated with manual judgments in the context of scientific summarization. I furthermore proposed an alternative and more effective approach for scientific summarization evaluation (Summarization Evaluation by Relevance Analysis - SERA). Results revealed that in general, the proposed evaluation metric achieves higher correlations with semi-manual pyramid evaluation scores in comparison with ROUGE.

Chapter 3

Text Categorization in the Health Domain

3.1 Introduction

In addition to the large amount of biomedical literature and the need for summarization, there is an increasing demand for use of electronic health records such as clinical notes and reports, as well as other forms of health-related textual data such as social media. The growing amount of health-related textual data requires non-manual processing for purposes such as improving health care, public health surveillance, quality measures, and improving wellbeing of individuals.

Text categorization is a key step in many Natural Language Processing (NLP) tasks to better organize, analyze, understand, and search data. With the raise of textual health-related data in recent decades, it has become challenging for healthcare professionals to utilize such data in an efficient way to address key problems in healthcare. Preventable medical errors have been shown to be a major cause of injury and death in the United States [85, 166, 267] and in fact the 3rd leading cause of death in the U.S. [166, 247], with an estimated incidence of 210,000 to 400,000 annual deaths [122, 166].

Mental health is an equally important health-related challenge that is sometimes unfortunately taken for granted. Mental health conditions are associated with impaired health-related quality of life and social functioning [231, 250]. Self-harm and

suicide, as serious mental health conditions, are among leading reasons of death worldwide [8, 192]. Each year an estimated number of 43,000 Americans die by suicide, on average there are 117 suicides per day, and about 500,000 people visit hospital for injuries due to self-harm [8, 35, 136].

In this chapter, my goal is to utilize NLP methods to take a few steps towards addressing some of these major challenges in healthcare. I first present a method for differentiating errors from insignificant stylistic variations in different versions of medical reports. I will then focus on categorizing and identifying levels of patient harm caused by medical or hospital errors reported in clinical narratives. Afterwards, I will switch focus to mental-health issues and discuss solutions that can help identify major mental-health problems such as depression and suicide through social media.

3.2 Identifying Critical Discrepancies in Clinical Reports

3.2.1 Background

In this section, our research goal is to identify the types of discrepancies between different versions of medical reports. In many hospitals, a key aspect in patient care and education of residents is the development of the necessary skills to interpret patient examinations and correctly report their findings. Reports are later examined by an experienced attending physician, who revises eventual interpretation errors or minor mistakes. Therefore, there might be discrepancies between these two versions of the report. Researchers have studied the frequency of the discrepancies in clinical reports [229, 265], as well as their impact on patient care [230]. In case substantive edits, *"significant discrepancies"* exist between the initial and the revised report. These discrepancies are due to potential medical error or mis-interpretation by the resident (for example in the case of mis-interpreting a radiology image). Prevention of such errors is essential to patient care and the education of the medical residents. On the other hand, *"non-significant discrepancies"* refer to cases where a report has been edited by the attending to only address reporting style and write-up issues. In Figure 3.1, examples of significant and non-significant discrepancies are shown (each example is a small section of a much longer report).

In recent years, systems to identify reports that have major discrepancies have been introduced. Sharpe, et al. [238] proposed an interactive dashboard that highlights the differences between reports written by residents alongside the version edited by attending radiologists. Kalaria and Filice [133] used the number of words differing between the preliminary and final report to measure the significance of the discrepancies. However, deviation detected using this measure does not fully capture the difference between reports with significant discrepancies and non-significant ones, as

	Significant discrepancies	Non-significant discrepancy
Preliminary report (resident radiologist)	"*No acute hemorrhage. No extra-axial fluid collections. ~~The differentiation of gray and white matter is normal~~.*"	"*Postsurgical changes related to right thoracotomy with surgical packing material and hemorrhagic blood products in the right lower chest.*"
Final report (attending radiologist)	"*Subtle hypodensities in the inferolateral left frontal lobe and anterolateral left temporal lobe likely represent acute cortical contusions. No acute hemorrhage. No extra-axial fluid collections. Small area of encephalomalacia in the right parietal lobe.*"	"*Postsurgical changes related to right thoracotomy with surgical packing material and large amount of hemorrhagic blood products in the right lower chest.*"

Figure 3.1: Example of significant and non-significant discrepancies between reports. The stroked-through text has been removed from the preliminary report by the attending radiologist, while the underlined sections have been added.

dissimilarities in the writing styles between residents and attending radiologists can also cause differences in word counts.

We propose an accurate and effective two-stage pipeline to distinguish between significant and non-significant discrepancies in radiology reports. In other words, given a set of preliminary radiology reports with the respective final reports, we identify those with significant discrepancies. The first stage of our pipeline employs an ontology of radiology terms and expressions to identify reports with no significant differences. The remaining reports are then separated by a Support Vector Machine (SVM) classifier. We evaluate the impact of a diverse set of textual, statistical, and assessment score features on the performance of the second-stage classifier. Some of these features have been previously used to assess the quality of the text summarization and machine translation systems. Results illustrate significant improvement over the baseline (up to +14.6% AUC, -52% FNR) and show the effectiveness of the proposed approach. Our focus on false negative rate is motivated by the fact that each missed significant

discrepancy is a missed opportunity to educate a resident about a significant error in interpreting an examination.

To summarize, the main contributions of this work are as follows: (*i*) We introduce an approach for automatically classifying the type of discrepancies between preliminary and final radiology reports. (*ii*) We explore the use of summarization and machine translation evaluation metrics as features identifying reports with significant discrepancies. (*iii*) We provide extensive evaluation of different aspects of the proposed pipeline.

3.2.2 Related work

A related–yet ultimately different–problem to the one studied in this Section is the classification of clinical reports based on their content. In this task, which falls under the text classification domain, the goal is to classify radiology reports into a discrete set of predefined categories. For example, Nguyen and Patrick [191] aimed at grouping radiology reports into cancerous or non-cancerous cases using an SVM. Chapman, et al. [40] presented a system for detecting reports with mediastinal findings associated with inhalational anthrax. Percha, et al. [210] classified reports by breast tissue decomposition using a rule based classification scheme. Johnson, et al. [127] proposed a hybrid approach that combines rules with SVM to classify radiology reports with respect to their findings. Bath, et al. [19] introduced a classifier to determine the appropriate radiology protocol among those available for each disease. Their semi-supervised system takes advantage of the UMLS[1] ontology.

Researchers have also proposed methods for quantifying or comparing the quality of text in various domains. For example, Louis and Nenkova [161] introduced a model for classifying sentences in news articles into general/specific depending on the level

[1] `https://www.nlm.nih.gov/research/umls/`

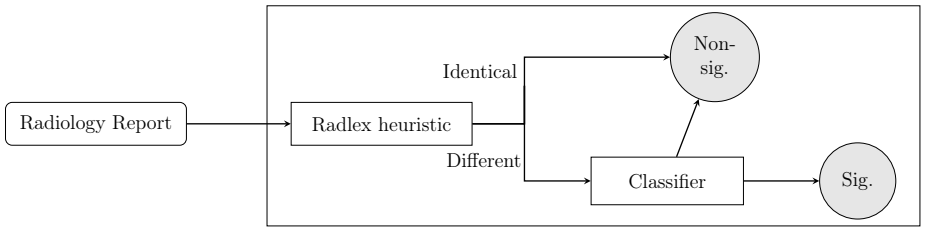


Figure 3.2: Overview of the proposed approach. The radiology reports are first classified by the Radlex heuristic. If there is no Radlex difference between a preliminary and the associated final report, the case is classified as non-significant discrepancy (*Non-sig* in the figure). Otherwise the case is sent to the a binary classifier for further analysis. The classifier which works based on several textual features, classifies the reports as having either significant (*Sig.* in the figure) or non-significant discrepancies

of the information carried by each sentence. Their classifier uses word, syntax, and language modeling features. Feng, et al. [95] explored a range of text features such as discourse properties, language modeling features, part-of-speech-based features, and syntactic features to quantify text complexity. Zeng-Treitler, et al. [275] proposed a system to grade the readability of health content; their tool employs lexical, syntactic, semantic and stylistic characteristics to accomplish such goal. Ashok, et al. [9] proposed an SVM classifier based on part of speech and lexical distributions, sentiment features, and grammatical properties to predict the success of novels. Lastly, Louise and Nenkova [162] proposed a model for predicting the appropriate length for a textual content in response to a specific information need.

Another line of related work is detecting plagiarism; systems designed for such task are concerned with determining if a given document was plagiarized from another source. To do so, current approaches in literature attempt to capture the significance of differences between a suspicious text and a source document (e.g., [1, 215, 246]).

Most of the previous efforts in plagiarism detection are centered on the retrieval aspect to find the original source of plagiarized content; thus, they focus on information and passage retrieval. Our problem differs from plagiarism detection in that our system takes as input a a candidate-source pair (preliminary and final reports) and attempts at classifying the significance of differences between them; instead, in plagiarism detection, the goal is the retrieval of source document.

3.2.3 Methodology

We propose a two stage pipeline for classification of type of discrepancies in radiology reports based on their significance. The overview of our approach is shown in Figure 3.2. In first stage, we utilize a heuristic based on domain ontology to identify non-significant discrepancies. In next stage, reports that are labeled as significant by the heuristic are processed by a classifier that exploits a variety of textual features. Specifically, we adapt features that are originally used to evaluate text summarization and machine translation systems to our problem. The following sections provide details about each one of these two stages.

3.2.3.1 Stage 1: Domain ontology

We first link the significance of the discrepancies to the differences between the domain specific concepts in the reports. To extract domain specific concepts, we use RadLex[2], which is a comprehensive ontology of radiology terms and expressions with about 68K entries.

The domain specific concepts between the preliminary report and the final report are then compared. There might be cases in which there are no difference between the concepts of radiology reports but in one report some concepts are negated.

[2] `http://www.rsna.org/radlex.aspx`

As an example, consider these two sentences: " *... hypodensities in the inferolateral left frontal lobe ...*" and "*... no hypodensity in the inferolateral left frontal lobe ...*". Although the radiology concepts are identical, the negation might indicate significant discrepancy. Therefore, we also consider the negations in which the RadLex concepts appear to prevent false classification. To detect negations, we use the dependency parse tree of the sentences and a set of seed negation words (*not* and *no*). That is, we mark a radiology concept as negated if these seed words are dependent on the concept. If the RadLex concepts of the reports are identical and the negations are consistent, we classify the type of changes as non-significant. We call this stage, the RadLex heuristic (As indicated in Figure 3.2). A more comprehensive negation detection algorithm (*NeGex* [39]) was also evaluated; however, its results did not show any significant improvement.

The RadLex heuristic highly correlates with human judgments in identifying non-significant changes, as shown in Section 3.2.4.2. However, this simple heuristic is not accurate for detecting the significant discrepancies. In other words, if RadLex terms or their associated negations are not consistent, one can not necessarily classify the report as significant.

3.2.3.2 Stage 2: Classification using textual features

To address the shortcoming of the RadLex heuristic, we propose a binary classifier. The classifier uses diverse sets of textual features that aim to capture significance of discrepancies in radiology reports. The features that we use include surface textual features, summarization evaluation metrics, machine translation evaluation metrics, and readability assessment scores. We briefly explain each of these feature sets and provide the intuition behind each one of them.

Surface textual features. Previous work used word count discrepancy as a measure for quantifying the differences between preliminary and final radiology reports [133]. We use an improved version of the aforementioned method as one of the baselines. That is, in addition to the word count differences, we also consider the character and sentence differences between the two reports as an indicator of significance of changes.

Summarization evaluation features. ROUGE[3] [155], one of the most widely used set of metrics in summarization evaluation, estimates the quality of a system generated summary by comparing it to a set of human generated summaries. ROUGE has been proposed as an alternative to manual evaluation of the quality of system generated summaries which can be a long and exhausting process. Rather than using ROUGE as evaluation metric, we exploit it as a feature for comparing the quality of the preliminary radiology report with respect to the final report. Higher ROUGE scores indicate that the discrepancies between the preliminary and the final reports are less significant. We utilize the following variants of ROUGE:

- ROUGE-N is the N-gram precision and recall between the preliminary and final report, where N is the gram length (e.g., N=1 indicates a single term, N=2 a word bigram, and so on.) We consider ROUGE-1 to ROUGE-4.
- ROUGE-L compares the two reports based on the Longest Common Subsequence (LCS). Intuitively, longer LCS between the preliminary and the final report shows that the quality of the two reports are closer and therefore differences between the two are less significant.
- ROUGE-S computes the skip-bigram co-occurrence statistics between the two reports. It is similar to ROUGE-2 except that it allows gaps between the bigrams.

[3]Recall-Oriented Understudy for Gisting Evaluation

Skip-grams are used in different NLP applications; they consider additional n-grams by skipping middle tokens. Applying skip-bigrams without any threshold on the distance between tokens often results in incorrect matches (e.g. we do not want to consider all "the the" skip-bigrams in a sentence with multiple "the" expressions). To prevent this, we limit the maximum allowed distance to 10 which is empirically chosen.

Machine translation evaluation features. The Machine Translation (MT) evaluation metrics quantify the quality of a system-generated translation against a given set of reference or gold translations. We consider the final report as the reference and evaluate the quality of the preliminary report with respect to it. Higher scores indicate a better quality of the preliminary report, showing that the discrepancies between the preliminary and final versions are less significant. In detail, we use the following MT metrics: BLEU [200], Word Error Rate and METEOR [83].

- BLEU (Bi-Lingual Evaluation Understudy): In our setting, BLEU is an n-gram based comparison metric for evaluating the quality of a candidate translation with respect to several reference translations. It is conceptually similar to ROUGE-N, except being precision-oriented. Specifically, BLEU combines a modified n-gram-based precision and a so-called "Brevity Penalty" (BP), which penalizes short sentences with respect to the reference. Here, we use the BLEU score of the preliminary report with respect to the final report as a feature that indicates the quality of the preliminary report.

- Word Error Rate (WER): WER is another commonly used metric for the evaluation of machine translation [241]. It is based on the minimum edit distance between the words of a candidate translation versus reference translations; we

consider WER as the following formula:

$$\text{WER} \stackrel{def}{=} (100 \times (S + I + D)/N)$$

where N is the total number of words in the preliminary report; S, I, and D are the number of Substitutions, Insertions, and Deletions made to the preliminary report to yield the final report.

- Metric for Evaluation of Translation with Explicit word Ordering (METEOR): METEOR is a metric for evaluation of machine translation that aligns the translations to the references. Here, we want to find the best alignment between the preliminary report and the final report. In addition to exact matches between terms, METEOR also accounts for synonyms and paraphrase matches between the words and sentences which are not captured by previous features such as ROUGE. We use both the WordNet[4] [174] synonyms and RadLex ontology synonyms for calculation of the METEOR score.

3.2.3.3 Readability assessment features.

To quantify complexity of textual content and the style of the reports, we use readability assessment features. Here, "style" refers to reporting style of the radiology reports, such as lexical and syntactic properties. Choice of vocabulary and phrases, length of the sentences, and structure of the sentences are examples of such properties. In detail, we use the Automated Readability Index (ARI) [140] and the Simple Measure Of Gobbledygook (SMOG) index [169]. These two metrics are based on distributional features such as the average number of syllables per word, the number of words per sentence, or binned word frequencies. In addition to these statistics, we

[4]WordNet is a large English lexicon containing synonym relations between the words.

Table 3.1: Agreement rate between the RadLex heuristic and two annotators A and B.

		RadLex	A	B
non-significant	RadLex	1.0	0.964	0.942
	A	0.964	1.0	0.906
	B	0.942	0.906	1.0
count=139	Fleiss κ = 0.880			
significant	RadLex	1.0	0.557	0.492
	A	0.557	1.0	0.934
	B	0.492	0.934	1.0
count=61	Fleiss κ = 0.468			

Agreement for significant and non-significant reports are separately presented. Both raw agreement rates as well as Fleiss κ between the annotators and the RadLex heuristic are shown.

also consider average phrase counts (noun, verb and prepositional phrases) among the features.

3.2.4 Empirical results

3.2.4.1 Experimental setup

We use a collection of radiology reports with discrepancies obtained from a large urban hospital for evaluation. These reports contain two main textual sections: *findings*, which contains the full interpretation of the radiology examination, and *impression*, which is a concise section that highlights important aspects of the report. We use both sections for evaluation of our proposed pipeline. We use 10 fold cross validation for evaluating the proposed classification scheme.

3.2.4.2 Classification using RadLex ontology.

As explained in Section 3.2.3, we first classify the reports using the RadLex ontology and the negation differences between the preliminary and final versions of the report. We ran this method on 200 randomly sampled reports from the dataset; two annotators were asked to label the reports based on significance of discrepancies. The annotators were allowed to label a case as "not-sure" if they could not confidently assign a label for the report. The agreement rates between the annotators and the RadLex heuristic is shown in Table 3.1. As illustrated, RadLex heuristic is highly correlated with human judgments and the Fleiss κ for non-significant reports is above 0.8, which can be interpreted as perfect agreement [105, 149]. However, the simple RadLex heuristic's performance for the reports labeled as significant is low. Thus, we conclude that RadLex concept differences between the reports do not necessarily indicate that the changes between them is significant. As we show in next section, the proposed classification scheme with the textual features can solve this problem for reports with RadLex differences.

3.2.4.3 Classification by textual features.

To evaluate our proposed classification approach, a radiologist manually identified types of discrepancies of 150 randomly sampled radiology reports that include RadLex concept differences.

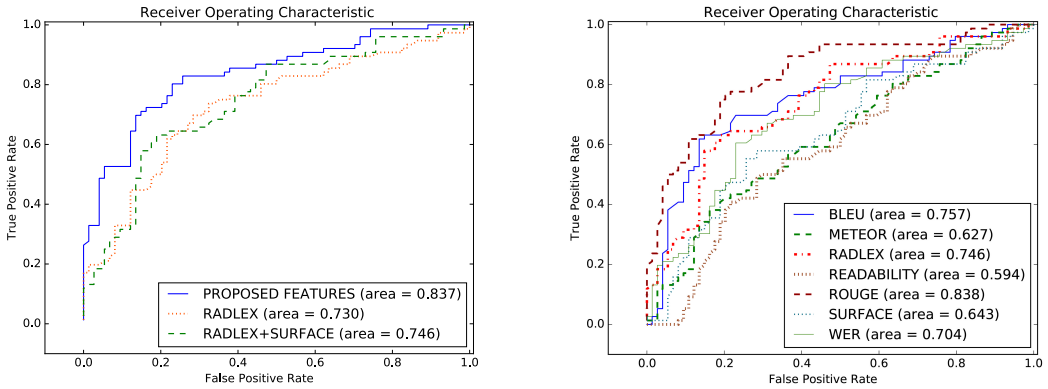
Feature analysis. Table 3.2 shows the cross validated classification results using the set of features described in Section 3.2.3. We use an SVM classifier with linear kernel. We report F-1 score and False Negative Rates (FNR) for significant reports, and the overall area under the curve and accuracy. We consider the following baselines: (i) Surface textual features including character, word and sentence differences

Table 3.2: Results of classifying significant reports.

Methods	F-1	FNR	AUC	ACC
Baselines				
Sf (Improved v. of [133])	0.650	0.329	0.642	0.633
RL	0.690	0.355	**0.746**	**0.707**
Sf+RL	**0.694**	**0.329**	0.730	0.700
Our methods				
Rd	0.568	0.421	0.594	0.553
BL	0.709	0.184*	0.757	0.660
M	0.604	0.368	0.627	0.580
Rg	0.767*	0.197*	0.838*	0.753*
Rg+BL	0.739*	0.237*	0.831*	0.727*
Rg+M	0.775*	0.184*	0.847*	0.760*
Rg+WER	0.702	0.211*	0.746	0.660
Rg+BL+M	0.780*	0.184*	**0.843***	0.767*
Rg+BL+M+RL	0.769*	0.211*	0.841*	0.760*
Rg+BL+M+RL+Rd	**0.797***	**0.171***	0.837*	**0.787***

The table shows the F-1 score (F1) and False Negative Rate (FNR) for significant reports as well as overall Area Under the Curve (AUC) and Accuracy (ACC) based on different set of features. The top part of the table shows the baselines and the bottom part shows our proposed features. Sf: Surface features – character, word and sentence differences; RL: RadLex concepts and their associated negation differences; Rd: Readability features; M: METEOR; BL: BLEU. Rg: ROUGE. Asterisk (*) shows statistically significant improvement over all baselines (two-tailed student t-test, $p < 0.05$).

between the reports (Indicated as "Sf" in the table). (ii) RadLex concepts and associated negation differences (Indicated as "RL"). (iii) Surface textual features along with RadLex concepts and negation differences (RL+Sf). Results based on different sets of features are presented. We experimented with all possible combinations of features; for the sake of brevity, we only report combination of features of significance.



(a) Comparison of the proposed pipeline with the baselines.

(b) Comparison of individual features.

Figure 3.3: ROC curves of the proposed method.

We observe that majority of the proposed features outperform the baseline significantly. One feature set performing worse than the baseline is the readability features. As described in Section 3.2.3.3, readability features mostly capture the differences between the reporting styles, as well as the readability of the written text. However, the reporting style and readability of the preliminary and final report might be similar although their content differs. For example, some important radiology concepts relating to a certain interpretation might be contradictory in the preliminary and final report while they both follow the same style. Thus, the readability features on their own are not able to capture significant discrepancies. However, when used with other features such as ROUGE, they are able to capture style differences that are not realized by other features especially in insignificant change category. This causes the performance of combined metrics to increase.

ROUGE features are able to significantly improve over the baseline. When we add METEOR features, we observe a further improvement over ROUGE alone. This

is likely due to the fact that METEOR considers synonyms in aligning the sentences as well, which is not captured by ROUGE. However, we note that METEOR by itself underperforms the baseline. We attribute this to the concept drift that may have been caused by consideration of synonyms in METEOR as observed in high False Negative Rate (FNR) of METEOR. The highest scores are achieved when we combine METEOR, ROUGE, BLEU, RadLex and readability features. We attribute the high performance of this setting to different aspects of reporting discrepancies captured by each of the features. ROC curve differences between our best performing features and the baseline (Figure 3.3a) further shows the effectiveness of our approach. Individual effects of features in terms of ROC curves are also compared in Figure 3.3b. As shown, ROUGE features are the most informative for identifying significant discrepancies.

Sections of the report. We evaluated which sections of the radiology report have more influence on the final significance of the discrepancies. As explained in Section 3.2.4.1, the reports have two main sections: *findings* and *impression*. As shown in table 3.3, *impression* section features have higher F-1 scores (+6.68%), lower false negative rates (-31.8%) and higher accuracy (+4.5%) than *findings* section. This is expected, since *impression* contains key points of the report. However, the best results are achieved when both sections are considered, thus indicating that the *findings* section contains valuable information that are not present in the *impression* section of the report.

3.2.4.4 Error Analysis

We examined the cases that our approach incorrectly classified. First, many of the false positive cases (i.e., reports that were incorrectly flagged as having significant discrepancies) were due to unnecessarily long length of preliminary reports. We saw

Table 3.3: Effect of sections.

Sections	F-1	FNR	AUC	ACC
Impression	0.772	0.197	0.821	0.760
Findings	0.725	0.289	0.817	0.727
All	0.797	0.171	0.837	0.787

Comparison of the results based on features extracted from different sections of the reports.

that in many cases, the preliminary report, especially in *impression* section, contains extra information that is later removed by the attending editor. In these cases, when almost half of the preliminary report is removed in the final version, our classification scheme fails to classify them as insignificant. According to the domain expert annotator, however, those removed sections do not convey any critical information. Since our features are mostly considering lexical overlaps between the reports, they fail to capture these special cases.

Second, we noticed that some of the false negative cases were due to only slight changes between the two reports. An example is illustrated below which shows a snippet from the preliminary and the final reports:

- **preliminary report**: "*Worsening airspace disease at the left base represents aspiration.*"
- **final report** "*Worsening airspace disease at the left base could represent aspiration.*"

This small change in the report is interpreted as a significant discrepancy between the two reports by the domain expert. Since there is only a slight change between

the two reports and the term *could* is not a domain specific term, our features fail to detect this case as significant. In this special case, the term *could* changes a specific interpretation from a definite fact to a possibility, should be considered as significant discrepancy.

Although the proposed approach misclassifies these cases, such discrepancies are very rare.

3.3 A Neural Attention Model for Identifying Harm in Clinical Narratives

3.3.1 Introduction

Preventable medical errors have been shown to be a major cause of injury and death in the United States [85, 166, 267]. To address these major concerns, healthcare systems have adopted reporting systems in clinical care to help track and trend hazards and errors in patient care [177, 267]. The data from these systems are later used to identify the causes of harm and actions that should be taken to prevent similar situations. These reporting systems allow frontline clinicians to report events that are relevant to patient care including both near misses and serious safety events. Near misses are events or situations where a hazard was identified before a patient could be harmed. For example, a wrong medication order that was never administered to a patient would be considered a near miss, hence reported as a no-harm event. Serious safety events on the other hand are situations where a patient was harmed. The above example would be considered a patient harm event if the nurse had actually administered the medication causing additional treatment, monitoring, or irreversible effects on the patient.

Although reporting systems have been implemented with the goal of improving patient safety and patient care, hospital staff are faced with many challenges in analyzing and understanding these reports [165, 177]. These reports which are narratives in natural language are generated by frontline staff and vary widely in content, structure, language used, and style. These reports include a textual field where the clinicians describe the safety event and its details in free-form text. While these texts provide valuable information about the safety event, it is challenging to perform large scale analysis of these narratives to identify important safety events. In this section,

we propose and evaluate Natural Language Processing (NLP) methods to identify cases that caused harm to the patient based on medical narratives.

One important aspect in patient care is to identify events that have contributed to or resulted in harm to the patient [267]. There have been many efforts in characterizing harm to the patients based on their severity. The most common is a harm categorical system that indicates the severity of the harm to the patient [5]. These categories range from an unsafe condition (which describes an event where there was no error, but had capacity to cause harm) to death (which is an event where an error has caused or contributed to the death of a patient). These harm categories are described in Table 3.4 in detail [5].

The patient incident narratives can be complex and it is challenging to identify the cases of harm from these reports. These reports often consist of multiple events. For example, consider a case where a patient is found on the floor in the emergency department (ED) with no physical signs of injury. This is initially entered as a no-harm case. However, later when the patient is transferred to the radiology for an x-ray as a precaution, a small fracture is discovered from the x-ray. Therefore, while the ED staff originally entered the event as a no-harm event, the radiology department would revise this as a harm event.

For these reasons, reporting harm is often miscategorized. While most events are eventually recategorized by a department manager or patient safety officers who have a more global perspective of events, this recategorization incurs additional time, resources, and expenses leading to missed opportunities to address the actual event in a timely fashion. We present a method for identifying the severity of harm from narratives regarding incidents in patient care. While there is a growing number of work in categorizing patient safety reports, none has looked at the modeling of general harm across all event types [96, 195]. Our method is based on a neural network

Table 3.4: Categories of errors in patient care.

Cat.	Description	Example
A	No error, capacity to cause error	Confusing equipment
B1	Error that did not reach the patient (due to chance)	Wrong medication label discovered
B2	Error that did not reach the patient (because of active recovery efforts by caregivers)	Mislabeled specimen in a laboratory discovered on a regular checking
C	Error that reached patient but unlikely to cause harm	Multivitamin was not ordered on admission
D	Error that reached the patient and could have necessitated monitoring and/or intervention to preclude harm	Regular release metoprolol was ordered for patient instead of extended-release
E	Error that contributed to or resulted in temporary harm	Blood pressure medication was inadvertently omitted from the orders
F	Error that could have caused temporary harm requiring initial or prolonged hospitalization	Anticoagulant, such as warfarin, was ordered daily when the patient takes it every other day
G	Error that resulted in permanent harm	Immunosuppressant medication was unintentionally ordered at wrong dose
H	Error that necessitated intervention to sustain life	Anticonvulsant therapy was inadvertently omitted
I	Error that contributed to or resulted in death	Beta-blocker was not reordered postoperatively

The categories are defined by Agency for Healthcare Research and Quality [5]. The severity of harm increases from top to bottom. Categories {E,F,G,H,I} are harm categories while {A,B1,B2,C,D} are no-harm events.

model consisting of several layers including a convolutional layer, a recurrent layer, and an attention mechanism to improve the performance of the recurrent layer. Our method is designed to capture local significant features as well as the interactions and dependencies between the features in long sequences. Traditional methods in general and domain specific NLP rely heavily on engineering a set of representative features for the task and utilizing external knowledge and resources. While these models have been shown to work reasonably well for different tasks, their success relies on the type of features that they utilize. Apart from the feature engineering efforts, these approaches usually model the problem with respect to certain selected features and ignore other indicators and signals that might improve prediction. In contrast, our approach only relies on the text in the patient incident narratives and it does not rely on any features or external resources, making it generalizable. Through extensive evaluation on two large datasets, we show that our proposed method is able to significantly outperform the existing approaches of identifying harm in clinical care. Effective identification of harm can help the hospital staff save time both during analysis and reporting. Furthermore, a more accurate and immediate classification of harm can also help to better prioritize resources to address safety incidents, which subsequently improves general patient care.

3.3.2 Related work

There has been a growing number of work in categorizing patient incident and safety narratives in clinical care. Fong et al. [96] explored both the unstructured free-text and structured data elements in safety reports to identify and rank similar events. They evaluated different search methods utilizing bag of words features, structured elements, and topic modeling features to rank and identify similar events. In another work, Ong et al. [195] explored the similar problem of identifying extreme-risk events

in patient safety and incident reports using Naive Bayes and SVM classifiers with bag of words features. In contrast to these works, we focus on the problem of identifying harm and categorizing the harm based on its severity in medical narratives. We present neural network methods that are able to capture information from the complex narratives regarding safety events without utilizing any external features. We compare our results with feature based methods and show that our proposed methods are significantly superior in identifying and classifying harm in patient incident reports.

The problem of identifying harm in patient safety reports is a type of text classification problem. Traditional approaches in text classification include methods to extract features from text and then use the feature vector as an input to a classifier such as SVM [6]. More recently, neural networks have shown success in many NLP tasks including text classification. Two of the more widely used neural network architectures have been Convolutional Neural Networks (CNN) [151] and Recurrent Neural Networks (RNN) [91]. Collobert et al. [59] were one of the first to utilize CNNs in many NLP tasks including text classification. In particular, they proposed a CNN architecture which operated on one-hot encodings of words; their model was based on the original CNN architecture of LeCun et al. [151] with adaptations to the NLP domain and showed improvements on several NLP tasks. Later CNNs were further explored for sentence modeling and classification tasks [86, 135, 139].

In the biomedical domain, there have been many efforts in classifying biomedical text and narratives based on different tasks. Many works have looked at the specific problem of indexing biomedical literature using MeSH[5] terms. These works have mostly used supervised learning frameworks with bag of words features, named-entities, and ontology specific features [274]. More recently, Rios and Kavuluru [223]

[5]Medical Subject Heading

utilized CNNs for this task; they showed that CNNs are more effective than feature based methods in biomedical indexing. Xu et al. [271] used a CNN architecture, with multiple sources of word embeddings and evaluated its effectiveness on the tasks of biomedical literature indexing and clinical note annotation. Our method, in contrast, is based on an extension of CNNs with recurrent layer as well as an attention model to improve performance on longer sequences. We compare our methods with a CNN baseline and show that our methods can significantly outperform the baselines. Our focus is on the challenging task of identifying harm in patient incident reports where the incident narratives are often complex, consisting of multiple chained events in a single narrative. Our proposed model, is designed to capture these complexities. Our initial results published in [55], also shows the effectiveness of this method in classifying safety reports into their respective categories.

3.3.3 Methods

We present a general neural network architecture for identifying harm in patient safety reports. As explained in §3.3.1, the narratives regarding patient safety can be complex and identifying harm to the patient is challenging in these reports. To be able to perform this task effectively, we will need a model that is able to capture both local features as well as the language usage in the entire report. To achieve this goal, we propose a neural network consisting of several layers where each layer is designed to address the aforementioned challenges. Our approach does not require feature engineering and it learns to identify significant features from the raw text automatically. We first describe the general outline of our model and then we describe each component in more detail.

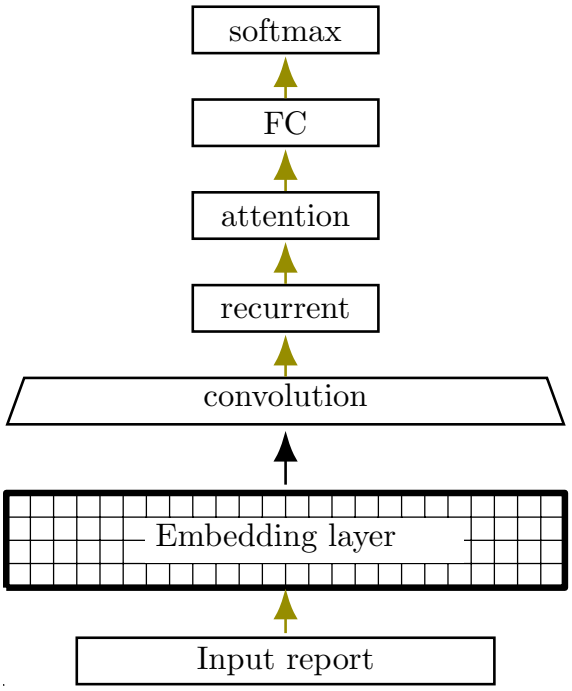


Figure 3.4: The outline of the proposed model. Input report consists of the raw text of the report, embedding layer does the pre-processing and represents the input as a matrix. FC=Fully Connected layer.

3.3.3.1 The outline of the model

The proposed architecture is shown in the Figure 3.4. The input report is first pre-processed and represented as a matrix corresponding to word embeddings. Word emdeddings or distributed representations of words aim to embed (represent) words with dense vectors such that words with similar properties have similar vectors [14]. These embeddings can be general and pre-trained or can be trained according to the task at hand. Then a convolutional layer extracts the significant local features that are helpful for identifying harm in the report. Next, a recurrent layer captures the interactions of the local features along the entire sequence of the words in the report.

In the next layer, we propose an attention model which serves to overcome the problem of recurrent networks in compressing an entire sequence in a single vector by focusing the attention to the important timesteps (steps of the sequence) in the recurrent layer. Finally, the output of the attention model is a vector which is passed to a fully connected layer and a softmax classifier identifies the level of harm associated with the report. We now explain each of these layers in detail.

3.3.3.2 Embedding layer

This layer pre-processes the raw text corresponding to the medical report and represents it as a matrix of real valued numbers. This matrix consists of embeddings of the words in the report. We tokenize the text using a simple white space tokenizer and we lowercase all the words. We then transform the input sequence of tokens into a sequence of dense distributional vectors. Specifically, given a sequence of tokens W where $W = \langle w_1, w_2, ..., w_n \rangle$ and w_i's are the input sequence tokens, the embedding layer represents each token w_i as a d dimensional vector x_i, and the sequence W will be represented as a matrix of real valued numbers $\mathbf{X}$ with dimensions of $\mathbf{X} \in \mathbb{R}^{(n_{\max} \times d)}$ where $n_{\max}$ is the maximum sequence length. Text inputs with length larger than $n_{\max}$ will be cropped and text inputs shorter than the $n_{\max}$ are padded with zeros. The value of $n_{\max}$ is determined empirically.

3.3.3.3 Convolutional layer

Convolutional layer is repsonsible for extracting local features from the input text. Convolutional Neural Networks (CNNs) [151] have been previously used in sentence modeling and classification tasks [135, 139]. A CNN is a neural network that consists of two main operations: convolution and pooling. A convolution is an operation between two functions f and g where f is the primary vector and g is the filter. The convolution

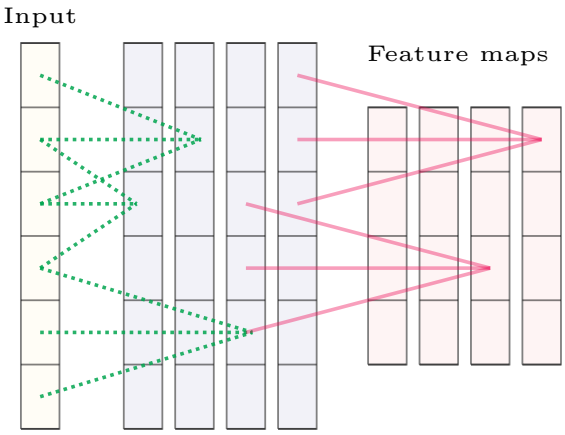


Figure 3.5: Convolutional layer. A convolutional layer takes a series of tokens as input and applies l filters of size k (green dotted arrows in the figure) to derive the feature values over a local window of tokens; $l = 4$ and $k = 3$ are shown here (all filters are the same size). To produce the component's output, a max pooling layer (red arrows in the figure) considers region sequences of length n and keeps the highest feature value for the sequence $n = 3$ is shown here.

operation between f and g, evaluated at entry n is represented as: $(f * g)[n] = \sum_{i=-K}^{K} f[n-i] \times g[i]$

Where $*$ denotes the convolution operation and $L = 2K - 1$ is the length of the filter. Here, f is the input to the convolution (word vectors obtained from the embedding layer).

Features are extracted by convolution of the input text with a number of linear filters, adding a bias term and applying a non-linearity. The result is called a feature map. The trained weights in these filters correspond to a linguistic feature detector that learns to recognize a specific class of n-grams where $L \leq n$. A max-pooling operation is used after the convolution to extract the significant features.

Multiple feature maps. Similar to convolutional networks for object recognition [151], we use multiple feature maps with different filters to capture various aspects of the input sequence. Figure 3.5 illustrates how the feature maps are constructed from the input. First the convolution and non-linearity are applied to the input and then the max pooling derives the resulting feature maps. The final output of this layer at each time-step is the concatenation of the feature maps at that time-step.

3.3.3.4 Recurrent layer

The result of the convolution layer is a sequence of vectors each of which is the concatenation of the feature maps at corresponding time-step. Convolutional layer is able to extract significant local features that are important for our task. However, the interactions between the words are not captured specially if the words are distant from each others. Recurrent Neural Networks (RNNs) are a family of neural networks that are designed to process a sequence of values. We use an RNN on top of the result of the convolution layer to capture interactions along the entire sequence of words. RNNs are an extension of multilayer perceptrons in which the output of each step is used as an additional input to the next step. Specifically, the activations arrive at the hidden layer of the network from both the current external input and the hidden layer activations one step back in time. The general formulation of an RNN is as follows:

$$h(t) = g(W^{(h)}h(t-1) + W^{(x)}x(t)) \tag{3.1a}$$

$$\hat{y}(t) = \text{softmax}(W^{(s)}h(t)) \tag{3.1b}$$

Where $h(t)$ shows the hidden state of the RNN in time step t, $x(t)$ is the input sequence at time step t, $W^{(h)}$, $W^{(x)}$, and $W^{(s)}$ are the weights associated with the hidden state, input, and softmax, respectively, and g is an activation function such as RELU [72].

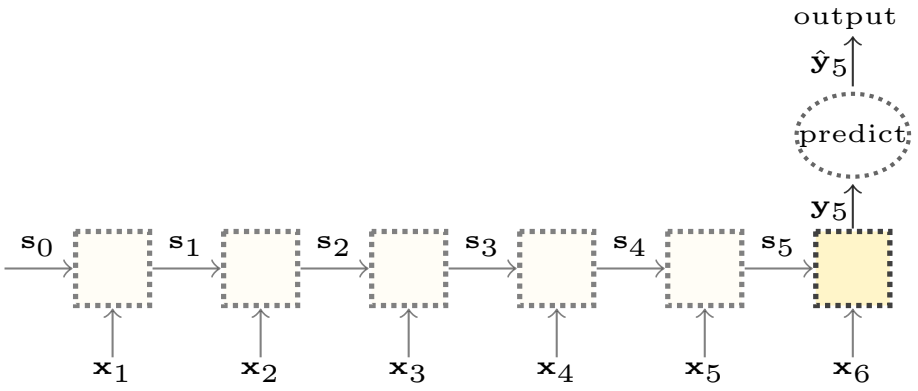


Figure 3.6: Representing a sequence with an RNN.

In sequence modeling tasks, the final hidden state of the network can represent the whole sequence and can be used for making predictions [99]. This final hidden state, in theory, can capture all the information in the entire sequence. This is because the output of each timestep is used as an input to the subsequent timestep in the network. Figure 3.6 illustrates the prediction made at the last hidden state of the network.

RNN variants. Training the general formulation of RNNs in practice is difficult due to the exploding and vanishing gradient problems (gradients becoming exceedingly high or become exceedingly close to 0 after only a few timesteps) [203]. For the exploding gradient problem, a common solution is to cap the gradient value at a specific maximum threshold. There has been some variants of RNNs that assist the gradient flow and mitigate the vanishing gradient problem. Most notable are the Long Short Term Memory (LSTM) [114] and Gated Recurrent Unit (GRU) [44].

LSTM adds additional gates to the regular hidden layer of a recurrent network to assist the gradient flow and allow the network to be effectively trained. These gates control the amount of information to be forgotten or preserved throughout the sequence. Concretely, we are referring to the formulation of Graves [103] for LSTM.

GRU proposed by Cho et al. [44] makes each recurrent unit to adaptively capture dependencies of different time scales and similar to LSTM, GRU also has gating units that control the flow of information through the computational graph. The difference with LSTM is that the GRU does not have a separate memory cell. We use the exact formulation of Cho et al. [44] for GRU.

Bidirectional RNNs. In order to also capture the backward dependencies and interactions between different parts of a sequence, a backward RNN is also trained which can encode the information from the future time steps [102, 234]. The hidden states of the backward RNN are then considered along with the corresponding hidden states of the forward RNN (e.g. by concatenation) at each time step and used in the subsequent layers.

Let $\boldsymbol{x} = \langle x_1, ..., x_n \rangle$ be the input to the Recurrent layer. Then the bidirectional RNN over the time steps $t = 1, ..., n$ is as follows:

$$h_t = \langle \overrightarrow{h_t}; \overleftarrow{h_t} \rangle \tag{3.2}$$

where "$\langle \cdot; \cdot \rangle$" shows the concatenation operation and $\overrightarrow{h_t}$ ($\overleftarrow{h_t}$) is the forward (backward) RNN defined as follows:

$$\overrightarrow{h_t} = \overrightarrow{RNN}(x_t); \quad \overleftarrow{h_t} = \overleftarrow{RNN}(x_t) \tag{3.3}$$

Where RNN$(\cdot)$ is the feed forward RNN cell in the general form, LSTM, or GRU.

3.3.3.5 Attention model

We use an attention model on top of our recurrent layer to be able to capture the local features that are more important in the task at hand. The limitation of using the regular recurrent network for the classification task is that the last time step of recurrent network loses some information about the sequence, specially when the sequence length becomes large [44]. This will not be a significant problem in short sentence classification tasks, but in our problem, the reports can have several sentences and the sequence length can be long. While in theory, the last step of the RNN is able to encode all the important information in the entire sequence, in practice it tends to focus more on the more recent time steps [253] and therefore loses some information specially about the earlier time steps. Using a bidirectional RNN can partially mitigate this problem where the last state of the backward RNN along with the last state of the forward RNN are able to capture the information in beginning and the end of the sequence. However, bidirectional RNNs are still suffering from the same information loss problem.

Inspired by recent work in machine translation [11] and document modeling [272], we propose to address this problem using a soft attention mechanism. Attention mechanism helps in constructing a context vector over the input that automatically incorporates the important parts of the input. The attention mechanism is shown in Figure 3.7. Particularly, instead of only considering the last hidden state of the RNN (h_n), the attention model attends to the important timesteps by introducing additional weights (α's):

$$c = \sum_{t=1}^{N} \alpha_t h_t \tag{3.4}$$

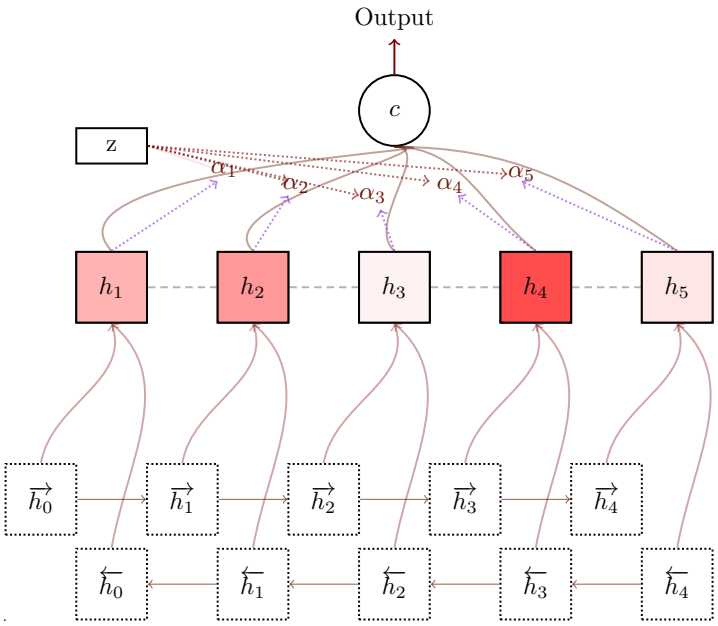


Figure 3.7: The attention mechanism over the recurrent layer. The states $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ show forward and backward hidden state of the RNN respectively. h_i's are the concatenation of the forward and backward RNN states. z is a context vector that attends to important time steps and α_i are the weights associated with each hidden state h_i. The figure shows an example where darker colors for the h_i show more importance in constructing c which is used as input to the next layer.

where t are the time steps in the input sequence and the weights α are learned according to the following softmax function:

$$\alpha_t = \text{softmax}(u_t^\intercal z) \tag{3.5}$$

where z is a context vector that helps in finding the weight importance of the local states h_t. This context vector can be seen as an input memory representation in memory networks and is jointly trained with the network. u_t is a feed forward function which we will define later and for a set of scores s_i, the softmax function returns a

probability distribution over the scores:

$$\text{softmax}(s_i) = \frac{\exp(s_i/\beta)}{\sum_j \exp(s_j/\beta)}; \tag{3.6}$$

with β being a parameter controlling the smoothness of the resulting distribution.

u_t in equation 3.5 is the result of applying a regular feed forward network over the hidden state h_t with weights U and biases b:

$$u_t = F(h_t) = \tanh(Uh_t + b) \tag{3.7}$$

Finally, our model at the top layer has a fully connected layer followed by a softmax non-linear layer that predicts the probability distribution of harm severity given an input report.

3.3.3.6 Training

Let Θ denote all the parameters in the network which includes the weights associated with each of the layers described in previous sections. The entire network is then trained to minimize the following loss function:

$$J(\Theta) = -\sum_{c=1}^{C} \mathbf{1}[y^* = c] \log \Pr(Y = c|\mathbf{x}) \tag{3.8}$$

where C is the number of harm severity classes and y* is the ground truth label for the input report $\mathbf{x}$, $\mathbf{1}[\cdot]$ is the indicator function, and the probability of each harm severity class is estimated through the network.

3.3.4 Experiments

3.3.4.1 Data

We use two large scale datasets consisting of patient safety and incidents reports sampled from various healthcare systems. These reports are sometimes referred to as

"patient safety reports" in the health informatics literature, but in general they are meant to identify and characterize errors in patient care.

Table 3.5: Dataset characteristics.

Statistic	Dataset 1	Dataset 2
Number of reports	28,539	248,213
Avg. report length (character)	411.4	239.2
Stdev report length (character)	370.9	187.2

Avg. refers to the average and std is the standard deviation.

Table 3.6: Distribution of harm levels across different severity categories.

	no-harm				harm
Dataset	A	B	C	D	[E-I]
1	39.3	13.8	19.9	13.1	13.9
2	11.7	12.9	40.4	31.6	3.4

Numbers show percentage of the entire data. The severity increases as we move from A to I. For the definition of the harm severity levels refer to Table 3.4 at page 2.

This study was approved by the MedStar Health Research Institute Institutional Review Board (protocol 2014-101). The characteristics of the datasets are outlined in Table 3.5. We observe that one of the datasets (DS2) is larger than the other one (DS1) and the length of reports are rather different between them. Each dataset consists of reports regarding different categories in patient care. The statistics about each of the harm levels are shown in Table 3.6. The harm events (right side of the Table) are usually much less frequent than the events with no actual harm (left side of the Table). We divide each dataset to 3 subsets of training, validation, and test with respective distribution of 60%, 20%, and 20% of the entire data. The hyper-parameters of the models were chosen empirically based on the performance on the validation set and the test set is preserved for evaluation.

3.3.4.2 Evaluation

We evaluate the effectiveness of our models in identifying harm by using standard classification evaluation metrics namely precision, recall, F-1 and Area Under the Curve (AUC).

3.3.4.3 Baselines

For comparing the performance of the proposed methods, we consider the following baselines:

- *SVM bow* - SVM with linear kernel with n-gram bag of words (bow) features [266]. We experiment with three types of features, n-grams of size {1}, {1,2}, and {1,2,3} (we respectively abbreviate the resulting models with *bow1*, *bow2*, and *bow3*).
- *MNB bow* - We also experiment with Multinomial Naive Bayes method for classification where Wang and Manning [266] show its effectiveness in many text classification tasks. We used scikit-learn (http://scikit-learn.org/) implementation of SVM and MNB.
- *CNN* - We consider CNN model for text classification which has shown good results in both general domain [135, 139] and biomedical domain [223].
- *LSTM* - We also compare against RNN (LSTM) classifier which is similar to the models used in [159, 255] (see Figure 3.6).

These methods form strong baselines with which we compare the performance of our models.

Table 3.7: The results of identifying harm vs no-harm events.

Method	Dataset 1				Dataset 2			
	P	R	F-1	AUC	P	R	F-1	AUC
Baselines								
SVM bow1	81.5	52.4	63.8	89.2	85.1	61.7	71.5	94.0
SVM bow2	81.8	55.7	66.3	89.7	84.8	64.9	73.5	94.6
SVM bow3	81.9	55.1	65.9	89.7	84.5	65.6	73.9	94.8
MNB bow1	81.1	52.8	64.0	83.0	66.6	73.9	70.1	89.3
MNB bow2	86.6	43.1	57.5	79.0	73.9	66.0	69.7	86.9
MNB bow3	**88.3**	37.7	52.9	77.2	77.0	60.0	67.5	85.1
CNN	75.7	63.9	69.3	90.4	80.2	67.2	73.1	94.7
LSTM	76.6	61.9	68.8	90.3	70.1	**75.9**	72.9	94.6
This work								
GRU CNN	75.8	68.3	71.8	91.0	77.9	73.2	75.5	95.0
Bi-GRU CNN	72.3	71.8	72.1	91.1	80.1	71.0	75.3	94.9
LSTM CNN	77.1	67.1	71.8	91.2	77.6	74.0	75.8	95.0
Bi-LSTM CNN	78.7	62.8	69.8	91.1	79.6	70.7	74.9	94.9
ATT GRU CNN	78.1	64.4	70.6	91.0	78.0	75.1	76.5	95.0
ATT Bi-GRU CNN	73.4	69.3	71.3	91.0	**87.3**	70.3	**77.9**	94.8
ATT LSTM CNN	69.4	**76.8**	**72.9**	**91.2**	78.8	74.9	76.8	95.0
ATT Bi-LSTM CNN	83.0	64.0	72.3	91.0	79.9	74.5	77.1	**95.2**

The top part of the Table shows the baselines while the bottom part shows the variants of the models presented in this work. Metrics are precision (P.), recall (R.), F-1 score for the harm category as well as the ROC Area Under the Curve (AUC). The numbers are percentages. On both datasets the F-1 scores of our attention models (starting with ATT) are statistically higher than than that of all the baselines (McNemar's test, $p<0.01$)

3.3.4.4 Model variants

We evaluate several variants of our models. The first variant is our model architecture which is the entire model presented in §3.3.3 minus the attention model. We consider two types of recurrent networks, GRU and LSTM, as well as their bidirectional variants (Bi-GRU and Bi-LSTM). We then evaluate our complete model which utilizes the attention mechanism. When considering attention, we evaluated both GRU and LSTM as the underlying recurrent layer. We abbreviate these models based on the layers from top to bottom. For example "*ATT GRU CNN*" corresponds to our attention model with GRU unit, while example "*ATT Bi-LSTM CNN*" corresponds to our attention model with bidirectional LSTM in the recurrent layer.

Design decisions and hyperparameters. We empirically made the following design choices and hyperparameter selection: We used embedding size of 100 for word vectors and we set the maximum sequence length to 100 words (smaller sequences are padded with zero vectors and larger sequences are cropped). For convolution, we used filters of length 2 to 5 with 128 channels each, max pooling of length 4, and merge the output of the filters by concatenation. For RNN, we use LSTM and GRU with hidden size of 100. We used dropout rate of 0.25 after convolution. Training was done with batch size of 128 and through 2 and 6 epochs for the larger and smaller datasets, respectively. Adam [141] was used as optimizer and early stopping was applied by monitoring accuracy on the validation set.

3.3.4.5 Results

We first consider the problem of identifying harm cases in the patient reports. That is, we classify a report as indicating some signs of harm to the patient (a harm case) or not (a no-harm case). The main results of our methods in identifying harm are

Table 3.8: The performance of our models in identifying fine grained harm categories.

	Temp./ perm. harm	Didn't reach pt.	Near Miss	Unsafe	Avg
Dataset1					
MNB bow2	41.4	39.7	67.7	77.0	64.1
SVM bow2	53.5	49.8	68.1	77.1	66.6
LSTM CNN	62.5	52.36	67.7	75.8	68.2
CNN	64.0	47.4	65.4	75.4	66.8
ATT Bi-LSTM CNN	64.1	51.6	68.5	76.1	68.7
ATT Bi-GRU CNN	64.5	48.9	69.7	77.0	**68.9**
Dataset2					
MNB bow2	43.4	52.0	82.8	51.8	71.0
SVM bow2	49.3	56.4	82.1	55.7	72.6
LSTM CNN	62.9	55.1	83.8	54.2	73.8
CNN	58.5	54.6	82.7	55.0	72.6
ATT Bi-LSTM CNN	59.3	59.2	83.8	54.6	74.1
ATT Bi-GRU CNN	66.6	60.2	83.4	59.3	**75.3**

Only the top model variants are shown. For the definition of the categories refer to §3.3.1 and Table 3.4. The numbers are F-1 scores percentages in each category.

illustrated in Table 3.7. The metrics are Precision, Recall, F-1 score for the harm category as well as the Area under the curve. We observe that our attention models (starting with ATT in the Table) are the best performing methods in both datasets evaluated by F-1 scores. In particular, the attention model using an LSTM recurrent unit (ATT LSTM CNN) achieves the highest F-1 of 72.9% on the first dataset and the attention model using a bidirectional GRU (ATT Bi-GRU CNN) achieves F-1 of 77.9% on the second dataset. While the results ranges are similar between the two datasets, in general we can see that the results on the second dataset are slightly higher. This is due to the datasets being generated at different healthcare systems and thus there are qualitative and quantitative difference between the datasets. As far as the baselines, we can see that in general, in terms of F-1 scores, traditional bag of words approaches [266] are not quite competitive. In terms of precision, the Multinommial Naive Bayes method using up to 3gram features (MNB bow3) achieves the highest overall scores on first dataset; however, its recall is very low, making it relatively ineffective. The SVM baselines work generally better on the second dataset compared with the first dataset, and they outperform the performance of CNN and LSTM baselines. For example, the best F-1 score on the second dataset is 73.9% which is for the SVM bow3 baseline. Our methods are still able to significantly improve over this baseline (compare the performance of *ATT Bi-GRU CNN* with *SVM bow 3*). Another trend that is worth noting is the significantly higher recall performance of our proposed models in comparison with the baselines. Recall is important in the task of harm detection, as any harm case can impact the patient and the method should minimize false negatives. We then compare the result of our method using a recurrent model on top of a convolutional model and observe how it can improve both the CNN and LSTM baselines. This suggests that while CNNs are effective in capturing the information in longer sequences, there is also some additional information that is captured when

Table 3.9: The results of our best method for the 1st dataset.

Category	F-1	P	R	Acc
Skin/Tissue	92.9	94.1	91.8	87.2
Surgery/Procedure	76.1	79.4	73	84.2
Restraints/Seclusion Injury	75	75	75	90.9
Airway Management	73.7	70	77.8	81
Blood Bank	71.4	100	55.6	98.1
Lines/Tubes/Drain	66	70	62.5	71.7
Medication/Fluid	63.9	77.5	54.4	93
Safety/Security	56.4	75.9	44.9	73.2
Diagnosis/Treatment	55	56.9	53.2	79.3
Miscellaneous	55	54.5	55.6	83.7
Fall	52.2	63.8	44.1	86.4
Diagnostic Imaging	50	44	57.9	83.7
Patient ID/Documentation	36.4	40	33.3	97
Lab/Specimen	0	0	0	97.2

We only show the results for common categories.

considering the interactions between the words along the entire sequence. We also observe that using a recurrent layer on top of a convoluational layer improves the performance (compare LSTM with our models in the Table), suggesting that local features captured by CNN are important in the final prediction.

Next, we evaluate the performance of our top models in fine-grain classification of harm severity on patients compared with the top baselines. Table 3.8 shows the performance on 4 levels of harm: Temporary or permenant harm, event that reached the patient but did not cause harm, near miss events, unsafe events. For description of these categories refer to Section 3.3.1 and Table 3.4. We observe that our method variant *ATT Bi-GRU CNN* achieves the best overall performance with average respective F-1 scores of 68.9% and 75.3% in datasets 1 and 2.

Table 3.10: The results of harm identification on the second dataset.

Category	F-1	P	R	Acc
Complication of P/T/T	81.8	80.6	83.1	85.7
Fall	79.7	86	74.3	95
Error in P/T/T	71.3	68.8	74	96.3
Miscellaneous	68.4	68.2	68.5	87
Skin Integrity	66.4	75.2	59.4	92.2
Equipment/Supplies/Devices	61.7	62.5	61	95.9
Transfusion	52.4	68.8	42.3	96
Medication error	49.4	69.4	38.3	98.3
Adverse Drug Reaction	46.2	61.5	37	80.8

Numbers are percentages. P/T/T refers to the category of Procedure/Treatment/Test

3.3.4.6 Analysis

To better evaluate the performance of our system and study the errors that it makes, we analyze the performance on each dataset based on each category of incident reports. The incident reports are categorized into several categories and there are often qualitative differences between the narratives in different categories. Tables 3.9 and 3.10 show the breakdown of results based on top common categories in dataset 1 and 2, respectively. We report the results of the best performing model variant on each dataset (i.e. *ATT LSTM CNN* for dataset 1 and *ATT Bi-GRU CNN* for dataset 2). On dataset 1 (Table 3.9) we observe that the model achieves very high scores in identifying harm in Skin/Tissue category with F-1 of 92.9% in identifying harm. Results on some other categories such as *Surgury/Procedure*, *Seclusion Injury*, *Airway Management*, and *Blood bank* are also relatively high. However, we observe that on some categories such as *Patient ID/Documentation* and *Lab/Specimen* the performance is low. We attribute the low performance in these categories to three

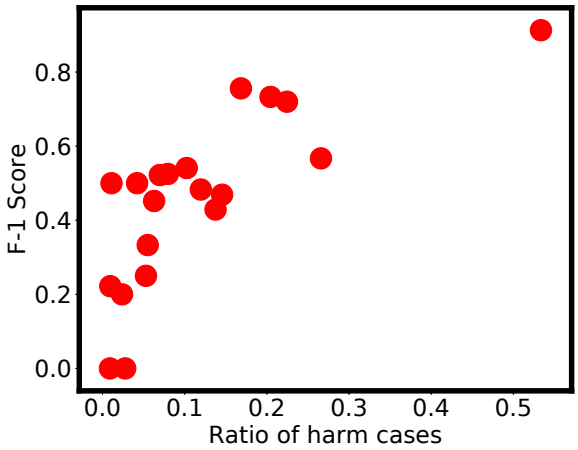


Figure 3.8: The performance of the our best method on dataset 1 based on each category. Each data point shows the results in a specific category as well as the ratio of harm cases in that category. The x-axis shows the ratio of harm cases.

main reasons: the total number of data in each category, the relative number of harm cases in the category, and the diversity of the type of reports in each category. We analyzed the distribution of the harm cases in each category. Some categories are more balanced in terms of harm and no-harm cases, while other categories are extremely unbalanced. We calculate the class ratios of harm in each category and compare the results based on these ratios. Figure 3.8 illustrates the performance of our method on each category and the ratio of harm cases in that category. Each data point shows the performance results in terms of F-1 based on the ratio of harm cases in that category. We observe that as the ratio of harm cases increases, the performance generally tends to increase. This is expected, as training the model on highly unbalanced datasets prevents the model to learn the appropriate weights associated with the positive class. The two categories at the bottom left side of Figure 3.8 are the categories with lowest results in Table 3.9. The respective ratio of harm cases in these categories are 0.009

and 0.027, while the ratio of harm cases in *Skin/Tissue* (the point on top right side of the Figure) is 0.53.

We also performed qualitative analysis on the reports in each category by inspecting the type of incidents in each category. We investigate the types of incidents in the best performing category in dataset 1 *Skin/Tissue* and most of the events are regarding pressure ulcer and wounds. On the other hand, looking at the *Lab/Specimen* category, there are many diverse types of errors and harm in this category such as collection issues, documentation problems, labeling issues, ordering issues, etc, that are very different in description, making it difficult for the model to learn all the nuances in this category. This reason, coupled with relative low number of harm cases in the dataset in this category, results in low performance. We believe that having more data would help improving the performance of the model.

3.4 Depression and Self-Harm Assessment through Social Media

In this section, I will switch focus to mental-health as another significant dimension of healthcare. Mental health is an increasingly important health-related challenge in society; mental health conditions are associated with impaired health-related quality of life and social functioning [231, 250]. I will particularly focus on depression, self-harm, and suicide which are among the most important aspects of mental-health. Self-harm and suicide, as serious mental health conditions, are leading reasons of death world-wide [8, 70, 192]. I will show how we can use attempt solving some of these major issues through processing language expressed in social media.

3.4.1 Background

Many individuals with mental-health conditions choose to express their problems and seek support and help through social media. The increasing ubiquity of social media makes it a ready and accessible platform for individuals who are at distress and willingly express their problems. Furthermore, given that these conditions are conventionally associated with high stigma, participating in discussions in anonymous or psedu-anonymous fashion makes it easy for these individuals to opt to using social media for receiving social support [128]. Therefore, social media has become a valuable platform for large-scale analysis of mental health data and this analyses can offer great insights into mental health. Generally, it has been shown that social media can have broad applicability for public health research as the data from social media can reflect a variety of characteristics about individuals [89, 202, 205].

In this section, we will focus on using Natural Language Processing methods to: *(i)* identify posts in mental-health forums that indicate signs of self-harm or suicide

to the user, and *(ii)* identify users with depression in general forums. This has clear scientific and clinical applications such as directing the attention of moderators to the identified critical posts in a timely manner.

3.4.2 Related work

3.4.2.1 Healthcare and Mental Health through Social Media

In recent years, healthcare has benefited enormously from social media data [87]. Many studies have investigated public health surveillance by utilizing the Twitter public data [41, 148, 202, 205, 206]. Results of these studies show consistency with other information resources for public health such as official reports released by governments, reports released by Centers for Disease Control and Prevention (CDC) and other online sources such as Google Flu Trends[6].

Social media has also become a popular platform for people with mental health conditions to express their feelings and seek support from other users. It has helped individuals with depression by providing them means to connect to people with shared experiences who can answer their questions and concerns [73, 194]. Consequently, the information from social media has become a significant resource providing more insight into psychological and mental conditions and problems. There is a growing body of related work analyzing mental health-related discourse and language usage in social media to better discover and understand mental health related concerns [7, 16, 64, 65, 66, 78, 178, 180, 222, 260]. De Choudhury et al. [77] explored social media to identify and diagnose depression among individuals. They analyzed the posting of a set of Twitter users through time and identified signals for characterizing the onset of depression in individuals. Park et al. [201] showed that depressed individuals perceived

[6] `https://www.google.org/flutrends/`

social media (Twitter) as a tool for social awareness and emotional interaction while non-depressed individuals are mostly regular information consumers. Schwartz et al. [235] used Facebook data to build a regression model to predict degree of depression in individuals. Portier et al. [214] conducted sentiment analysis on the cancer survivor forum content and compared the sentiment change of the user content before and after interaction with the community. There exist many other works on analysis of social media for mental health problems such as depressive disorders [75, 260], addiction [182], insomnia [123], schizophrenia [178] and various other conditions [67].

While many of the aforementioned mental health disorders are closely related to depression and suicidal behaviors, our focus in this section is to identify the severity of the content based on indication of self-harm risk to individuals and identify depression through general language usage. Our depression detection models only rely on text expressed in user posts and they are not dependent on any external or domain-specific features. Existing self-reported diagnosis detection datasets contain a limited number of both control users and diagnosed users. Thus to support training and evaluating neural models, we also constructed a substantially larger and more realistic dataset with over 9,000 depressed users matched with more than 100,000 control users[7].

3.4.2.2 Social Media and Suicide

Previous work has studied self-harm and suicidal behavior through NLP. Some researchers explored the language usage in content relating to suicide to identify signals of this behavior to predict suicidal actions. Thompson et al. [259] predicted the risk of suicide in military personnel and veterans using the clinical notes and online social media data (Facebook posts). They used a model based on Random Forest

[7]Our dataset is available upon request at: `http://ir.cs.georgetown.edu/data/reddit_depression/`

classifier [28] with bag-of-words features. Jones and Bennell [130] developed statistical prediction rules to discriminate between genuine and simulated suicide notes. Lester [152] analyzed the language of suicide notes to better understand suicidal behaviors in individuals. Coppersmith et al. [69] examined data from Twitter users who have attempted to take their life and provided an exploratory analysis of patterns in language around their attempt. Some researchers have analyzed suicidal behaviors through detecting sentiment and emotional variations of the content [43, 84, 211]. Prior work has also explored classification of suicidal content. Burnap et al. [30] proposed an ensemble classification approach to classify tweets into suicide related topics such as suicidal ideation, reporting of a suicide, memorial, campaigning and support. Braithwaite et al. [26] conducted a user study on a group of individuals and analyzed their Twitter posts using Decision Tree classifier to differentiate individuals with higher suicide risks from individuals who are not at risk. Finally De Choudhury et al. [79] proposed that social media could be used to predict shifts from mental health discussions to expression of suicide thoughts. Specifically, they analyzed language in Reddit[8]mental health community and employed a framework based on propensity score matching [226] to predict suicidal shifts in users. Unlike these works, our focus is triaging the content severity in mental health online forums based on the risk of self-harm to the users.

Recently, research on NLP methods for suicide detection was further motivated by a shared task of the 2016 Computational Linguistics and Clinical Psychology Workshop [115] on automatic identification of content severity in mental health forums. Most of the proposed methods, generally used Support Vector Machine (SVM) classifiers [71] or an ensemble of some other standard classifiers for identifying the content severity. We briefly describe the top 3 approaches: Kim et al. [138] used a

[8]https://www.reddit.com/

Stochastic Gradient Decent classification framework. They utilized the body of the text as the main source for feature extraction and represented the post by weighted TF-IDF[9]unigrams and distributed representation of documents [150]. Malmasi et al. [167] used a hierarchical classification framework. They employed a Random Forest meta-classification approach on top of a set of base classifiers. Finally, Brew [29] used SVM with Radial Basis Function (RBF) kernel; they utilized TF-IDF unigram and bigram features, author type, post information and position of the post in the thread as the features for the classifier.

In contrast to these works, our approach is feature-rich; many features that we use are not present in the aforementioned prior work, such as psycholinguistic, contextual, topic modeling and skip thought features (see the Methods section for details). We also utilize an ensemble classifier using different subsets of features. Our proposed models outperform the state-of-the-art by large margins.

While the aforementioned works only focus on triaging the content severity, we further utilize the triaging model to perform large-scale analysis of user interactions in this forum to gain insight on the impact of the forum on the users with mental health issues. We analyze the moderators' response time to users and show that without an accurate and efficient content triaging system, manually identifying severe posts in forums with large number of users is indeed difficult.

3.4.3 Suicide and self-harm risk assessment

Specialized online forums are a type of social media which are essentially communities in which users with common special interests engage in discussion. Mental health forums are one type of these specialized forums which are centered around users who have directly or indirectly been involved in mental health conditions. General social

[9]Term Frequency - Inverse Document Frequency

media platforms such as Twitter and Facebook in comparison with specialized forums are less topic-centric and more general purpose, in the sense that millions of users use them to discuss mundane events in their lives. While the signals coming from general social systems such as Twitter and Facebook are subtle and not directly about mental health, they are relevant and they have been previously utilized to support certain important tasks (e.g. [67, 235, 260]). On the other hand, online forums are specifically designed for discussion around specific topics and they attract users with similar interests and goals [74]. Users in general social media such as Twitter can choose to be pseudonymous or anonymous. On the other hand, to protect their users, many online mental health forums such as ReachOut[10] specifically enforce maintaining anonymous profiles. The moderators in many of these forums actively redact any post that could reveal the identity of a user. Such support for anonymity further encourages users to engage in sensitive mental health discussions and express their real thoughts and feelings. We first focus on these specific online mental health forums as anonymous support platforms centered around people with similar experiences and problems.

There are three stages that lead to suicidal action among individuals who are in some sort of mental distress [79, 239]: 1- thinking, 2- ambivalence and 3- decision making. In the first two stages the individual is experiencing thoughts of distress, hopelessness, and low self-esteem. In the decision making stage, the individual might show explicit plans of taking their life. Individuals might seek support in any of these stages and online health forums are a ready platform enabling these individuals to ask for support. In many online mental health forums, there are moderators or more senior members who help the users with mental distress. Troubled users who are at risk of self-harm need to be attended to as quickly as possible to prevent a potential self-harm act. However, the volume of newly posted content each day makes it difficult for

[10]https://forums.au.reachout.com/

the moderators to locate and respond to more critical posts. Effective online manual triaging of all the forum contents is highly costly and not scalable.

We propose an approach for automated triaging of the severity of user content in online forums based on indication of self-harm thoughts. Triaging the content severity makes it possible for moderators to identify critical posts and help a troubled user in a timely manner to hopefully reduce the risk of self-harm to the user. We propose a feature-rich supervised classification framework that takes advantage of various types of features in the forums. The features include lexical, psycholinguistic, contextual, topic modeling, and dense representation features. We evaluate our approach on data provided by ReachOut[1], a large mental health forum. We show that our approach can effectively identify the critical content which will assist the moderators in attending to the in-need users in a timely manner. We show that without an automatic way for identifying critical posts, the moderator's response time does not correlate with the severity of the posts, which further confirms that manually identifying these posts is a challenge for moderators. Finally, analysis of the user content on this forum shows that on average, the content severity of users tends to decline as they interact with the forum which is evidenced by the transition from more critical to less critical content.

3.4.3.1 Severity risk assessment

Our main objective is to determine the severity of the mental health forum posts based on signs of self-harm thoughts in the content. Triaging content severity enables moderators to attend to severe cases in a timely manner and hopefully prevent a potential self-harm attempt.

[1]www.ReachOut.com

Our approach for triaging the content severity is a supervised learning framework. In the following, we first define the severity categories, then we explain the features that we use for the classification and finally, we describe the learning algorithm.

Severity categories. We consider the following 4 levels of severity for the post content, as defined by [176]:

- **Green -** posts that do not show any signs or discussions about self-harm and thus do not require direct input from the moderators. These posts are usually general statements or follow up discussions that do not reflect any major concern.
- **Amber -** posts that include minor clues that might indicate signs of struggle by the user. These posts need the moderator's attention at some point, but prompt intervention is not necessary.
- **Red -** posts indicating that the user is in acute distress and moderators should attend to them as soon as possible.
- **Crisis -** posts indicating that the user is in imminent risk of self-harm. These posts could be about the authors themselves or someone that the author of the post knows. Moderators should prioritize these cases above all others.

Table 3.11 shows synthesized examples of posts in each of these severity categories[11]. Following the terminology used by Milne et al. [176], we consider the union of CRISIS, RED and AMBER categories as FLAGGED posts, because they indicate that user might be at risk and needs attention at some point. Similarly, we consider the union of two more critical categories, i.e CRISIS and RED as URGENT.

[11]The provided examples throughout this paper are very similar to the ones in the ReachOut forum. According to the data collection policies on protecting users' identities, we are unable to include the exact posts from the forum.

Table 3.11: Example of posts in each severity category.

GREEN	AMBER	RED	CRISIS
I'm proud that I was able to call and keep up a phone conversation with my mum.	There are so many stuff I'm thinking about, but my medications are slowing my thoughts down and making it more manageable	I feel helpless and things seem pointless. I hate feeling so downâĂę	I'm having some strong thoughts about ending my life, nothing helps.

Due to large volume of posts produced each day, it is not possible for moderators to identify all the critical posts in a timely manner. Our goal is to predict the severity of the forum posts' content so that the moderators can locate critical cases and attend to them as soon as possible. We propose a feature-rich machine learning approach utilizing psycholinguistic, topic modeling and contextual features.

Features. Since the forum posts are written in unstructured raw text, we extract representative features from the text that are helpful for the supervised learning. Particularly, we extract the following categories of features:

• **Bag of words** An standard approach for text representation is to model the text with bag of its constituent words. This results in a sparse vector for each text in which each element associates with a word in the vocabulary and is weighted according to some weighting scheme. We use the unigram and bigram bag of words representation of text with frequency of terms as their weights. Throughout the paper, when we refer to some textual content (e.g. post body) as features, we are essentially referring to the unigram and bigram bag of words representation of that text, unless otherwise

noted. Before representing the text with bag of words features, we perform standard minimal preprocessing on it by lowercasing and removing stopwords.

• **Psycholinguistic** The psycholinguistic features are meant to capture the different dimensions of a user's mental state through analysis of their language usage.

– *LIWC*: Linguistic Inquiry and Word Count (LIWC) [208] is a tool that captures quantitative data regarding various psychological dimensions given the user's textual writings. It utilizes several psychological lexicons along with a text analysis module that associates text with different psychologically-relevant categories. We use this tool to extract different psychological attributes from the language expressed in the users' posts. While LIWC provides over 100 distinct attributes, our experimentation showed that the affective attributes, drive attributes, tonality, informal language usage, anxiety attributes and negation are the most helpful for this task.

– *Emotions*: Emotions are very closely related to suicide. Therefore, the emotion that is reflected by the post can be a good indicator about level of severity of the content. For example, if a user's post indicates the "anger" emotion, it is more likely to be severe in comparison with a post that shows the "happiness" emotion. To quantify the emotions associated with a specific post, we use DepecheMood [244], a lexicon with emotional probabilities associated with more than 37000 terms. The emotions considered by the lexicon are "fear", "amusement", "anger", "annoy", "apathy", "happiness", "inspiration" and "sadness". To obtain the overall distribution of emotion over these categories for a post, we average the emotion distribution of all words in the post to obtain probability of each emotion given the post. We use these probabilities as features for the classification. In addition to the specific probabilities, we also consider the dominant emotion of the post as a separate feature.

– *Subjectivity*: Similarly, subjective posts are more likely to be related to a severe post than an objective post. We utilize the MPQA subjectivity lexicon [268] to differentiate between the subjective and objective posts. This lexicon contains contextual subjectivity about words or phrases that indicates expression of an emotion, opinion, stance, etc.

• **Contextual** One characteristic of online forums is that they are designed to support user discussion. Therefore, having information about the context of a given post in the discussion thread provides additional information about its content. We extract the following contextual features:

– *Author's prior posts*: Author's prior posts in the thread captures the development of thoughts by the user and also in combination with the body of the post captures whether the post deviates from the author's prior posts in a significant way.

– *Prior discussion*: The posts preceding a target post and written by other users help in capturing surrounding discussion and development of thoughts for the target user. Specifically, we consider a window of 3 posts by other users preceding the target post as the context of the post in the thread. Limiting the window size to 3 is due to our observation that in long threads, the discussion usually deviates after a few posts, hence considering all the posts would introduce noise to the model[12]. We could also consider the posts succeeding the target post as additional features, however, that would not correspond to a real-world scenario. In a realistic setting, the goal is to triage the content on the forum as soon as they are posted and therefore, to comply with this setting, we do not consider any features relating to content submitted after the target post.

[12]We experimented with context window of sizes 1 to 5. The best performance was for context size of 3, therefore we chose window of 3 posts as the context size.

– *Last sentence*: Finally, some critical posts are long, and mostly about some mundane and usual events that happen; in these posts, there is a sudden change at the end of the post indicating that the user might be at risk. Take the following example which is a snippet from the beginning and ending part of a longer post (Parts indicated with [...] are omitted for brevity):

"Now, I think we all know what it's like to be rejected by friends, dates, etc. While I have been stood up by a certain friend a few times, this really got to me. My dad said on tuesday [...]
... I woke up today and I since morning just don't know what to do anymore. I feel like I have nothing to live for and nothing makes me happy anymore."

In this example, most of the body of the post does not indicate any immediate risk to the user. However, this sudden change in the user's mental state shows that this content is potentially a severe case. If we only rely on the features capturing the entire post, the mental state shift will not be apparent as most of the post do not show any signs of risk. Therefore, we also consider the last sentence as a separate feature; we utilize the LIWC attributes for the last sentence to focus on the final mental state of the user and to eliminate some of the dilution that may occur in longer posts.

• **Topic modeling** We use the abstract "topics" that occur in the collections of posts as another set of features for classification. Topic modeling [21] is a widely used approach for discovering the latent semantic structures ("topics") in a text body. Latent Dirichlet Allocation (LDA) [22] is a generative model that describes how the documents in a dataset are created. A brief description of the LDA generative process is as follows:

1. For each document:
   (a) Draw a distribution over topics

(b) Generate each word in the document by:

i. Drawing a topic β_j according to the distribution selected in step (a).

ii. Drawing one word from the V words in the topic β_j

Using this generative process, the LDA model tries to find a set of topics that are likely to have generated the collection. We trained the LDA topic model on the entire forum posts to obtain the latent topics associated with each post and we used these topics as additional features[13].

• **Skip thought vectors** Bag of words representation of the post is a sparse representation in which most of the entries are zero. More recently, approaches have been proposed for obtaining a dense representation of sentences that can encode syntactic and semantic properties of sentences in vectors. Skip thought vectors [144] are one such model that use "sequence to sequence" models on pairs of consecutive sentences to learn the sentence encoding. Their model consist of a encoder-decoder framework in which the encoder maps words to a sentence vector and a decoder is used to generate the surrounding sentences. By analysis through several tasks, Kiros et al. [144] showed that this approach results in good sentence encodings when trained on a sufficiently large corpus. We use this model to encode the forum posts in dense representations. We average the vector representation of all sentences in the post to encode the entire post.

• **Forum metadata** Forum metadata such as number of post views, length of the thread, and number of post "kudos", a ReachOut feature similar to "likes" on Facebook, are additional features that we considered. Motivated by previous research that identified the time of day of online activity as a useful mental health signal

[13]We limited the number of topics to 100. We experimented with 20,50,100, and 200 topics and 100 topics was the optimal choice.

[64, 76], we also consider the broad temporal categories (day and night) as well as more fine-grained intervals (morning, afternoon, evening, and night). However, we did not observe an increase in the classifier's performance with the addition of the temporal metadata attributes.

Learning algorithm. After extracting features, we use supervised multi-class classification for triaging the user posts into different severity categories. We use the XGBoost Tree Boosting [42] as the learning algorithm. We experimented with several other standard classifiers such as logistic regression, random forest, and SVM, but XGBoost showed the best results.

Let the dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ consist of n different training instances in which the i th instance is represented by a feature vector $\mathbf{x}_i$ and label y_i. In matrix notation, the entire feature vector and the labels are represented as $(\mathbf{X}, \mathbf{y})$. Given this dataset D, the XGBoost tree ensemble model uses an ensemble of K additive functions (regression trees) to predict the output $\hat{y}_i$:

$$\hat{y}_i = \phi(\mathbf{x}_i) = \sum_{k=1}^{K} f_k(\mathbf{x}_i), f_k \in \mathcal{F} \tag{3.9}$$

where ϕ represents the model that predicts the output given the feature vector $\mathbf{x}_i$, $\mathcal{F}$ is the space of all regression trees, and K is the total number of regression trees used. The essential part of the model is regression trees f_i. To learn f, given the model output $\hat{\mathbf{y}}$ and the true class labels $\mathbf{y}$, the following regularized objective function is optimized over the training data:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3.10}$$

where l is a differentiable convex loss function (e.g. squared loss $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$), and $\Omega(f_k)$ is the regularizing function that penalizes the complexity of the

Table 3.12: Distribution of the labeled forum posts in the dataset.

	Train set		Test set		Total	
Severity Category	# posts	% posts	# posts	% posts	# posts	% posts
CRISIS	39	4	1	0	40	3
RED	110	12	27	11	137	12
AMBER	249	26	47	19	296	25
GREEN	549	58	166	69	715	60
Total	947	100	241	100	1188	100

Percentages are rounded.

functions to prevent overfitting. The model is trained additively by greedily adding f_k that most improves the model based on equation 3.10. The additive function f_k is also learned by a greedy tree growth algorithm. Several approximations are used that can quickly optimize the objective function. For more details on these steps, refer to the XGBoost reference [42].

In addition to the single classification model, we also utilize the ensemble of several XGBoost classifiers, each trained on a different subset of features from the entire feature space. We empirically determine the optimal subsets of features. By ensembling, we use multiple classifiers to obtain better performance than individual classifiers. Intuitively, we take advantage of several conceptually different models (each of which obtained by training on a different feature set), and we aggregate their predictions to obtain the final class label. We use the majority voting ensembling approach which has been shown to balance out the weaknesses of individual classifiers [147, 196].

Formally, let $\{\phi^{(1)}, ..., \phi^{(m)}\}$ be m models obtained by training the classifier on m different feature sets $\{\mathbf{X}^{(1)}, ... ,\mathbf{X}^{(m)}\}$. Similarly let $\{\hat{\mathbf{y}}^{(1)}, ..., \hat{\mathbf{y}}^{(m)}\}$ represent the output predicted by models $\{\phi^{(1)}, ..., \phi^{(m)}\}$. For the i th instance in the dataset,

the majority voting ensembling approach predicts the class label $\hat{y}_i$ according to the following:

$$\hat{y}_i = \underset{c \in \{c_1,...,c_T\}}{\text{argmax}} \left(\left| \{ j \in \{1,...,m\} : \hat{y}_i^{(j)} = c \} \right| \right) \tag{3.11}$$

where $\{c_1, ..., c_T\}$ is the set of all possible class labels.

XGBoost has several hyperparameters including the learning rate (η), the minimum sum of the weigths of all observations in a child (*min-weight*), and the maximum depth of the tree (*max-depth*). We used the default parameters which are $\eta = 0.3$, *min-weight* $= 1$ and *max-depth* $= 6$. We did not observe any performance gain by modifying the default recommended hyperparameters.

3.4.3.2 Experiments

Data. The data that we use in this research are forum posts from ReachOut.com which is a very large and popular mental health forum in Australia and receives about 1.8 million annual visits [173]. While this forum provides a discussion platform for ordinary topics such as life, family and friendship, its main purpose is to support discussions around more critical topics such as addiction, sexuality, identity and mental health problems. Most of the users and visitors are young people aging between 14 to 25 years old. ReachOut employs several senior moderators as well as younger people who volunteer for forum moderation. These moderators focus on cases that require attention and try to help these individuals by engaging in the discussion, showing compassion and support, and providing links and resources to the individuals.

We use a subset of the ReachOut forum containing 65,755 posts, 1,188 of which had been labeled by moderators based on 4 different categories of severity. The posts were annotated by three experts who achieved a FleissâĂŹs Kappa of 0.706 and pairwise CohenâĂŹs Kappa scores ranging between 0.674 and 0.761 which shows

substantial agreement [264]. The dataset contains separate training and testing sets; its characteristics are outlined in Table 3.12. The posts occurred between July 2012 and June 2015, with labeled posts being from May 2015 to June 2015. The posts were written by 1,647 unique authors. Each post contains several fields such as the post date and time, username of the author, number of kudos, subject of the thread, and the textual body of the post.

Data collection and privacy. The full details of the data collection and the discussion on the ethical issues are discussed by Milne et al. [176]. While analysis of the mental health forum data provides many benefits, there are always trade-offs between the benefits and the risk to the privacy of the individuals. Milne et al. [176] identified three groups of participants to whom the data collection and annotation process could cause harm: to the researchers who annotated the data, to the researchers who accessed the data, and to the people who authored the content. The data collection process ensured that the researchers were aware of the distressing nature of the content. To protect its users, forum members of the ReachOut are instructed to keep themselves safe and anonymous. Furthermore, the moderators in the forum actively redact any content that might reveal the identity of the users. The organizers further protected the forum member's anonymity by restricting researchers in contacting the individuals in the forum, distributing the data, and cross-referencing individuals against other social media.

Evaluation. Following Milne et al. [176], we use the accuracy and F-1 scores for evaluating the classification performance to be able to directly compare the performance of our approach with the state-of-the-art. To aggregate the scores for the individual categories, Milne et al. [176] used the macro average of F-scores for the

Table 3.13: Results of triaging content severity.

Methods	Macro Average over non-GREEN categories		FLAGGED vs. GREEN		URGENT vs non-URGENT	
	F1	Acc	F1	Acc	F1	Acc
Baseline	31	78	75	86	38	89
Cohan et al. [54]	41	80	81	87	67	92
Brew [29]	42	79	78	85	69	93
Malmasi et al. [167]	42	83	87	91	64	93
Kim et al. [138]	42	85	85	91	62	91
This work (Single model)	47.2	93.9	90.0	91.7	73.1	92.9
This work (Ensemble model)	**50.5**	**94.7**	**92.2**	**93.4**	**75.5**	**94.6**

(a) Results on the test set.

Methods	Macro Average over non-GREEN categories		FLAGGED vs. GREEN		URGENT vs non-URGENT	
	F1	Acc	F1	Acc	F1	Acc
Baseline	29.0	87.4	78.2	80.6	64.2	86.7
This work (single model)	43.0†	89.6†	85.1†	86.1†	**78.3**†	90.8†
This work (ensemble model)	**44.5** ‡	**90.6** ‡	**88.1** ‡	**88.8** ‡	77.6†	**91.4**†

(b) Results on the training set

Numbers are percentages. FLAGGED category is AMBER ∪ RED ∪ CRISIS. URGENT category is RED ∪ CRISIS. F1 is F1-Score and Acc is Accuracy. Baseline is the SVM classifier on post body (unigram and bigram features). Table (a) presents classification results and comparison with the baseline and state-of-the-art on the test set. Table (b) shows classification results on training set based on 10-fold stratified cross validation. For Table (b), †(‡) shows statistically significant improvement over the baseline (all other methods in the Table) according to the Student's t-test ($p < 0.02$).

Table 3.14: Features in our single and ensemble models.

Model	Features
Single model	Post body, forum metadata, subjectivity, emotions, contextual features, last sentence, topic modeling, LIWC
Ensemble model	1- Post body, forum metadata, subjectivity, emotion 2- Post body, contextual features, emotion features, LIWC 3- Post body, contextual features, last sentence 4- Post body, last sentence, emotion, sentiment 5- Post body, contextual features, topic modeling 6- Post body, contextual features, LIWC, clue words, forum metadata

The ensemble model is comprised of 6 classifiers with fewer number of features.

non-GREEN (critical) categories as the official metric for the CLPsych 2016 shared task. This metric emphasizes the importance of triaging among the critical categories. They also consider the F-1 and accuracy scores for binary classification of FLAGGED (i.e. CRISIS ∪ RED ∪ AMBER) vs. GREEN, and URGENT (i.e. CRISIS ∪ RED) vs. non-URGENT categories to capture the performance of systems in identifying critical posts. We also use these additional metrics to further evaluate the performance of our approach. FLAGGED classification shows that the post contains content indicating risk of self-harm to the user while URGENT indicates that the user is at a more imminent risk and needs prompt attention (see the Method Section for complete definitions of severity categories).

Baselines and comparison. We compare our methods with the top 4 performing systems among 16 total participating teams in the CLPsych 2016 shared task. To

better evaluate our methods, we also consider a simple baseline which is SVM classifier with unigram and bigram bag-of-words features extracted from the body of the post.

3.4.3.3 Results and analysis

The results of our models for triaging the content severity compared with the baseline and state of the art systems is presented in Table 3.13; it includes results on the test set 3.13a, as well as stratified cross-validation[14] results on the training set 3.13b. For prior work, we report the official results that are percentages without any precision points. The single model indicates the performance of our proposed model using a single classifier while the ensemble model is a model based on 6 different classifiers. The features used in each of the models are presented in Table 3.14. In the Analysis Section, we will discuss the effect of different features on the performance. As illustrated in Table 3.13a, our models outperform the baseline and all top performing state of the art systems by large margins. We observe that the non-GREEN macro average F1 score for the individual and ensemble models improves over the best system [138] by +12% and +17%, respectively. Similarly, we observe that the F1 scores for the FLAGGED category is 3% and 5% higher than the best system with the individual and ensemble models, respectively. Finally, in URGENT category, the individual and ensemble models achieve 73.1% and 75.1% F1 scores respectively, which shows large improvement over the state of the art. We observe similar improvements in the cross-validation results on the training set (Table 3.13b). Since we have 10 different folds on the training set, we also perform a statistical significance test and we observe statistically significant improvement over the baseline for both the single and ensemble methods (Student's t-test); the ensemble method also outperforms the single method

[14]The stratified cross validation in contrast to the regular cross validation preserves the distribution of the classes when splitting the data into train and test sets.

Table 3.15: Fine-grained classification results for each severity category.

Methods	Severity categories			
	CRISIS (1)	RED (27)	AMBER (47)	GREEN (69)
Baseline	0	39	53	90
Cohan et al. [54]	0	59	64	90
Brew [29]	0	65	61	88
Malmasi et al. [167]	0	58	69	93
Kim et al. [138]	0	65	61	94
Single model	0	67.7	67.4	93.7
Ensemble model	0	**75.5**	**76.1**	**95.2**

(a)

Methods	CRISIS	RED	AMBER	GREEN
Baseline	5.3	31.5	50.7	85.5
This work (single model)	17.0†	53.0 †	63.2 †	89.0 †
This work (ensemble model)	**21.3** ‡	**55.3** ‡	**69.1** ‡	**91.1** †

(b)

The numbers show macro-average F-1 scores in percentages. Last two rows show models proposed in this work. The top table (a) shows classification results and comparison with the baseline and state of the art based on each severity category on the test set. The numbers in parenthesis in front of each category is the total number of instances in that category. Note that CRISIS has only 1 instance and no system was able to detect that. Table (b) shows classification results by severity category on the training set (10-fold stratified cross validation). For Table (b), †(‡) shows statistically significant improvement over the baseline (all other methods in the Table) according to the Student's t-test ($p < 0.02$).

statistically in virtually all metrics. In particular, the single and ensemble models achieve 48% and 53% improvements over the baseline based on non-GREEN macro average F1 scores.

Table 3.15 shows the breakdown of results by each category. We present results on the test set in Table 3.15a and cross validation results on training set in Table 3.15b. It should be noted that there was only 1 CRISIS case in the test set and no team out of

16 teams were able to correctly identify this case. While our models were also unable to find the single CRISIS case, they show improvements over the state of the art in other categories. Specifically, we observe that the ensemble model achieves F1 score of 75.5% in RED which improves over the best performance (65%) by 16%. Similarly, we observe large improvement of F1 for the AMBER category (10%). Finally, our model also slightly improves upon the state of the art on the GREEN category. We also report results on the training set evaluated by 10 fold stratified cross validation (Table 3.15b). As illustrated, our methods achieve statistically significant improvement over the baseline in all severity categories. The overall lower performance on the CRISIS category is mainly due to the limited training data in this category. As shown in Table 3.12, there are only 40 CRISIS posts in the training set which is not enough for a supervised learning model to accurately estimate the optimal parameters.

Overall, the results show that both our single and ensemble models can effectively identify posts with critical content (FLAGGED) with F1 and accuracy of 92% and 93%, respectively, on the test set, providing large improvements over the state of the art.

In the rest of this section, we first analyze the effect of different features that we proposed to use for triaging the content severity. Then we analyze the types of errors that our model makes to better understand the robustness of our proposed approach. Finally, using the proposed triaging model, we investigate the potential effect of the mental health forum on the individuals.

Feature Analysis. In the "Severity Triaging" Section, we presented our proposed features for the task of triaging the content severity. Table 3.16 shows the effect of each of the features when added to the classification model. We do not show the combinations of features that perform significantly worse than the body of the text.

As illustrated, we observe that most of the proposed features have a positive effect on the performance of the system with the exception of skip thought vectors. The bag of words features of the body of the text achieve F1 score of 34.8% on the test set. Adding contextual features (prior posts by other users and user's previous posts in the thread) improves the results to 38.5%. Similarly, we observe that addition of forum metadata features (length, kudos, and post views), subjectivity and emotion features, and features from the last sentence also improve the performance. Topic modeling yields further boost to the performance of the system which indicates the effectiveness of latent topics inferred from the forum posts using the LDA model. We observe that LIWC features by themselves do not improve the results as much as topic modeling, however when combined with topic modeling features, greatest improvement is achieved (47.2% F1). This row (indicated by $*$) comprises all features in the single model reported in tables 3.13 and 3.15.

We build an ensemble of distinct models each of which trained on a different feature set. We experimented with various ensembles of the features. Last row of Table 3.16 shows the performance of the best ensemble model. We do not report other ensembles that resulted in suboptimal performance. The ensemble model that obtains the best results is comprised of 6 different feature sets outlined in Table 3.14. As evidenced by Table 3.16, each of these sets are helpful features that can capture different characteristics of the associated forum post; therefore when combined by ensembling, the weaknesses of single set of features on some instances are compensated by the others. Therefore, as the results show, the ensemble model is more effective in comparison with the single models.

We note that skip thought vectors (second row in Table 3.16) did not improve over the baseline. We also experimented with encoding the prior posts and authors posts with skip thought vectors but we did not observe any improvements. As shown by

Table 3.16: Effect of each set of features on triaging based on the test set.

	Macro average over non-GREEN categories			
Features	Acc	F1	P	R
baseline (body)	87.6	34.8	33.5	36.6
skip thought	87.5	33.5	33.4	34.1
body+contextual	90.3	38.5	36.5	40.8
+meta+subj	90.5	38.8	36.5	41.6
+lexical clues	90.9	40.2	38.3	41.3
+last sentence	92.3	42.8	43.0	42.8
+emotion	92.7	44.1	44.6	44.0
+topic	92.9	45.8	45.5	46.2
−topic+LIWC	91.8	41.9	41.7	42.6
+topic (∗)	93.9	47.2	48.9	45.8
Ensemble model	94.7	50.5	51.6	49.5

Numbers show percentages of macro averaged results for the FLAGGED categories (CRISIS ∪ RED ∪ AMBER). Acc: Accuracy, F1: F1-score, P: Precision, R: Recall. Body is the textual body of the post; "skip thought" is dense representation of text using skip thought vectors, "meta": forum metadata features; "subj": subjectivity features; "topic": Topic modeling features extracted using LDA, "LIWC": Linguistic Inquiry and Word Count features. Plus (+) signs show that the feature is added to the features in the above row and minus (−) signs show that the feature is eliminated from the above row. The row shown with (∗) indicates the features (listed in Table 3.14) used in the single model in tables 3.13 and 3.15. Accordingly, the last row is the ensemble model.

Kiros et al. [144], when trained on a sufficiently large data, skip thought vectors encode text in dense vectors that can capture underlying semantic and syntactic properties of the text; and thus useful to be used as features for classification. However, in this task we observe that classification using skip thought vectors does not result in any improvements. The lack of improvement by these vectors indicates that the vectors are not able to capture any information beyond what is provided by other features. This could be due to averaging the sentence vectors. We represent the post which consists of several sentences by averaging the vectors corresponding to each sentence; some of the information of the individual sentences might be lost when averaged with other sentences. Therefore, a better approach for composing the post vectors of its constituent sentence vectors could lead to better results.

Error Analysis. Error analysis shows that misclassification of content severity is mainly due to the following reasons:

1. Brevity of the posts and lack of sufficient background context.

Some URGENT categories that were misclassified are associated with a rather short post from which limited information can be obtained. For example, the following post is taken from a long discussion thread and is labeled as GREEN by the classifier while the actual label is RED.

"I got the reply from x about my complaint. All they did was make excuses for themselves. no help at all."

This post on its own does not show any risk to the user. However, reading the entire associated thread in the forum reveals that the author of the post had experienced a problem with their counseling service for their mental distress, and they were in need for mental help and support. To infer this context about this specific post,

the immediate surrounding posts are not sufficient and one needs to read the entire conversation.

In the model, we already consider the immediate surrounding posts as the context for the post. However, this may not capture the context in very long discussion threads (such as the above example). When we increased the number of previous posts to be considered as the context, we observed an overall suboptimal performance. This is because, generally, in long threads the discussion tends to change after a few posts. Thus, considering longer window of posts in a thread as context for a target post might result in adding posts that are not necessarily relevant to the target post and consequently introduce noise to the model.

2. Variations in tone.

In some misclassification cases, we observe sudden changes and variations in the tone of the post expressed by the user and that makes it difficult for the learning algorithm to correctly classify the associated severity. For instance consider the following post:

" I went to my favorite show last week and it was amazing. I usually feel very low, specially at nights. This was one of the rare times that I was actually happy for some time... Five days ago at school one classmate of mine bullies me and he shouts that he wishes me dead. I ignored him completely at the moment and I was totally fine. But when I got back home I felt like a total loser and the bad thoughts about myself started coming back."

In this post we observe that the user starts with a positive tone and then it changes to negative. Then the tone switches between positive and negative multiple times. This specific example is an AMBER case and the classifier mislabeled it as RED.

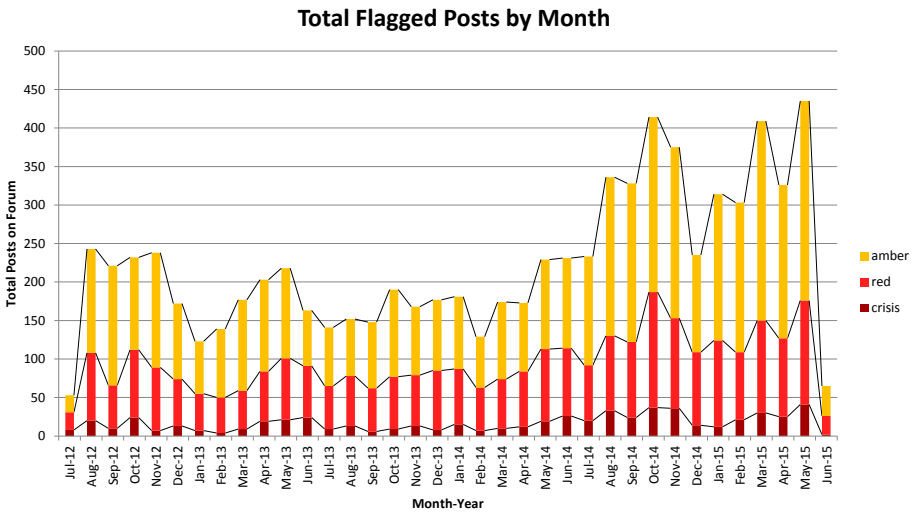


Figure 3.9: Volume of flagged posts on the forum.

Table 3.17: The number of FLAGGED users or URGENT users.

	First post		
Last post	FLAGGED	GREEN	Total
FLAGGED	93	37	127
GREEN	105	220	325
Total	198	254	

(a) FLAGGED

	First Post		
Last Post	URGENT	non-URGENT	Total
URGENT	30	16	46
non-URGENT	126	280	406
Total	156	296	

(b) URGENT

The table shows the number of users by the FLAGGED a or URGENT b post severity of their first post and their last post. Numbers in cells show the number of users whose first and last post severity corresponds to the associated column and row, respectively. For example 105 in table FLAGGED a corresponds to the number of users whose first post was FLAGGED and last post was GREEN.

Table 3.18: The number of users by average severity.

	First month		
Last month	FLAGGED	GREEN	Total
FLAGGED	120	46	166
GREEN	78	208	286
Total	198	254	

(a) FLAGGED

	First month		
Last month	URGENT	non-URGENT	Total
URGENT	40	31	71
non-URGENT	64	317	381
Total	104	348	

(b) URGENT

Numbers in cells show the number of users with average post severity in the first and last month corresponding to the associated column and row. For example 46 in Table (a) corresponds to the number of users whose average post severity in first month was GREEN and last month was FLAGGED.

In the proposed triaging model, we capture the user's final state of the mind by considering features from the last sentence. However, when there are too many tone variations in the post, the exact severity of the post might be misclassified. We note that the size of the training dataset was limited and therefore capturing these subtle cases requires more of similar training instances. Future work could investigate whether these variability of various psychological variables (e.g. tone) can be considered as a risk factor for individuals.

3. Long posts with only a small part containing concerning content.

In a few long posts, we observe only a small part showing signs of distress to the user, while the rest of the post has a neutral to positive tone. A misclassified example

with actual label of RED is shown below (Parts indicated with [...] are omitted for brevity):

"This book series is a roller coaster. Maze runner series, I'm onto the prequel book now. They are amazing [...] I've always been too resilient. I just hate everything and it confuses me. Maybe I'm tired of all this and want to do something.. I just... nothing is set. Yesterday Lora called and we talked like a lot about school, friends [...] It feels good to say, or type, all this."

This snippet is from a much longer post and as it can be observed, only the underlined part contains content that indicate mental distress to the user.

In such posts, the effect of the small negative part of the post is played down by the larger dominant neutral tone and therefore the model could mispredict this. In this case although still correctly identified as critical, the classifier misclassifies the severity level as AMBER instead of RED.

Overall, most of classification errors occur within the FLAGGED category; there are very few cases in the FLAGGED posts that are missed by the classifier and labeled as GREEN. This can also be observed in Table 3.13 in FLAGGED category performance which obtains F1 and accuracy scores of 92.2% and 93.4%, respectively. Our results are encouraging since they show that the model can effectively capture FLAGGED posts, i.e. all posts that indicate some signs of harm to the user.

User Analysis. We study the user content severity in the forum over time to analyze if it is helpful to the individuals. For the purposes of user analysis, we mostly rely on the binary classification of URGENT (CRISIS and RED) vs. non-URGENT, and FLAGGED (CRISIS ∪ RED ∪ AMBER) vs. GREEN categories. In these categories, as shown in Tables 3.13 (a and b), the ensemble classification model obtains F-1 scores

of 90% and 75% respectively (accuracy of 91% and 93%) and thus it is relatively reliable for studying larger scale trends of content severity in the entire forum. Figure 3.9 shows the results of severity triaging throughout all the posts in the dataset. As illustrated, there is a steady increase in the amount of FLAGGED posts. Given this trend, we examine patterns of post severity to understand the effects that the forum might have on the individuals. Specifically, we investigate the following research questions:

Q-1. **Does engaging with the forum have a positive effect on the users?**

Our analysis indicated a decline in the average content severity over time, which may indicate a positive effect of the forum on its users. However, further controlled trials should be conducted to carefully ascertain the causal nature of this relationship.

The dataset includes posts from the forum in a time window of 36 months during which we quantify the behavior of users. To measure the relation of user interaction with the forum, we split the users into two groups. Users are considered active if they have posted for two or more months on the forum, and inactive if they had only posted during a single month. We only consider active users for the analysis because for inactive users, the activity period of one month is too short to present a significant relation. In these 36 months, there are a total of 452 active users and 1,195 inactive users. We analyze the severity of the first post and last posts of users, average post severity during their first and last months of activity and finally, the trend lines of severity during entire time of interaction with the forum.

Tables 3.17a and 3.17b show the number by the severity of their first and last posts on the forum. A Chi-square test on the contingency tables was performed to ensure that the difference between the cells are interpretable. For both table 3.17a and 3.17b we found significant interaction, $\chi^2 = 58.4$, $p < .001$ and $\chi^2 = 21.4$,

Table 3.19: Statistics of r values.

	Avg.	Std dev.
Positive trend	0.68	0.34
Negative trend	-0.72	0.32

The average (Avg.) and standard deviation (Std dev.) of the r values of the trend lines for the positive and negative trends.

$p < .001$, respectively. In general, we observe that the users' last posts tend to be of lower severity than their first post. 81% of users whose first post received an URGENT label had a final post with a non-URGENT label. Only 10% of users whose first post was non-URGENT had a final post of URGENT. In both the FLAGGED and URGENT matrices, there were more users whose final posts was GREEN or non-URGENT than users who had FLAGGED or URGENT first posts.

Tables 3.18a and 3.18b show the comparison of the average user content severity in the first and last month of users' activity in the forum (Chi-square test showed that the results are interpretable with a significant difference of χ^2=86.47 ($p < 0.001$) and χ^2=52.82 ($p < 0.001$) for Tables 3.18a and 3.18b, respectively.). We observe a similar positive trend in the URGENT category; in the FLAGGED category, the number of users whose average initial content and last content is FLAGGED (120 users) is more than those whose content is shifted from FLAGGED to GREEN (78 users). However, there are very few GREEN users whose content eventually turned FLAGGED (46 users). Furthermore, the total number of users with first month FLAGGED posts (198) is higher than number of users with last month FLAGGED posts (166). These results also indicate that users' last posts tend less severe than their first posts.

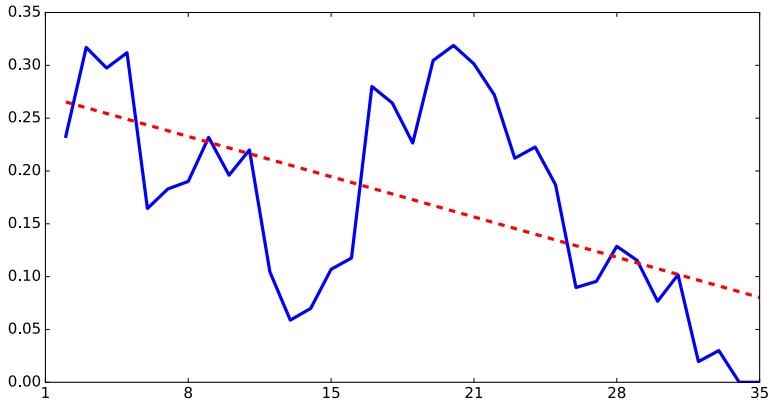


Figure 3.10: Example trend line of user post severity over time. x axis shows the month of activity. y is the average content severity in the month.

Table 3.20: Analysis of trend lines of severity over time for active users.

	FLAGGED vs GREEN				Fine-grained severity			
Threshold	Avg.	Stdev.	# positive	# negative	Avg.	Stdev.	# positive	# negative
0.02	-0.096	0.370	90	113	-0.068	0.236	76	120
0.05	-0.134	0.430	60	86	-0.104	0.285	43	84
0.10	-0.177	0.458	41	72	-0.129	0.320	32	65
0.15	-0.224	0.510	30	57	-0.159	0.350	23	53
None	-0.044	0.221	167	272	-0.032	0.151	153	298

FLAGGED vs GREEN indicates the trend change between FLAGGED and GREEN categories while fine-grained severity is for all 4 severity categories. *Avg.* shows the average of the slope of the trendlines. *Stdev.* is the standard deviation of the slope of the trendlines. *#positive* shows the number of users with positive slope of trendline. *#negative* shows the number of users with negative slope of trendline. Negative (positive) slope of trend line shows decreased (increased) content severity of the user over time. Threshold is used to filter out the effect of the flat trendlines; the considered trend lines in each row have an absolute value of slope greater than the value of the threshold in that row. Overall, the Table indicates that the content severity for majority of the users with non-flat trend line has decreased over time.

We believe this is because many users join this type of forums to get immediate support for a moment of crisis or acute mental distress. After some time, this initial distress is decreased, as reflected in the patterns of post severity. That is why the initial activity of users in general tend to be more severe than their final posts. We believe this could be for the following reasons: *(i)* Pattern of post severity drops off once the user is in a more stable mental state compared with their initial state of crisis. *(ii)* Interaction with the forum and engaging in discussion with other forum users might have resulted in reducing the acute distress in users (verifying the exact causal relation requires further user level controlled trials).

In addition to first and last months of activity, we also analyze the trends throughout the entire time of user activity. To do so, we consider the average severity of the posts in each month as a data point for that month, and we then fit a trend line to the data points. We consider the following numeric values for each category to be able to quantify the average severity in each month: CRISIS = 1.0, RED = 0.66, AMBER = 0.33, GREEN = 0.0. Using these numeric equivalent of severity classes, for each user, we associate an average severity for all their posts in each month. Then, we fit a linear model on this data to show the trend line of the content severity over time. Figure 3.10 shows a sample plot of the post severity for a user over time and its associated trend line.

To fit an appropriate trend line to the data, we minimize the squared error between the target trend line and the actual severity data points. Specifically, the equation of a trend line for variable x is given by $p(x) = m.x + b$ where m and b are the slope and intercept of the line, respectively. A negative (positive) trend line slope indicates that overall, the severity of user content has declined (increased). Given D severity data points $\{(x_i, y_i)\}_{i=1}^{D}$, the values of m and b are found by minimizing the squared

Table 3.21: Time in hours.

	First Month	
Last Month	FLAGGED	GREEN
FLAGGED	3.47	5.15
GREEN	3.62	7.02

(a) FLAGGED

	First Month	
Last Month	URGENT	Non-URGENT
URGENT	3.14	5.11
Non-URGENT	3.28	6.48

(b) URGENT

The average number of months the users stayed active in the forum based on the average severity of their content in the first and last months of activity.

error over the data:

$$E = \sum_{i=0}^{D} |p(x_i) - y_i|^2 \tag{3.12}$$

To check if a linear model is a applicable for our case, we calculated the r values associated with the trendlines. In particular, for each user we calculated the r value of their content severity trend lines and we calculated the average and the standard deviation of these values (Table 3.19). As illustrated, the average of r values are 0.68

Table 3.22: Average response time when a moderator was the first to respond.

Moderator Response Time					
	Total	Number	Percentage	Average Time	Stdev Time
CRISIS	608	147	24.18%	4.21	5.71
RED	2798	931	33.27%	4.53	6.17
AMBER	4642	1435	28.05%	4.46	6.60
GREEN	57707	892	1.55%	3.76	6.07
URGENT	3406	1078	37.96%	4.37	5.94
FLAGGED	8048	2513	37.88%	4.40	6.16

(-0.72) for the positive (negative) trends which is around 0.7 (-0.7). Absolute r values greater than 0.5 indicate high to strong linear relationship in the data [254]. Thus, linear trend analysis is a reasonable fit to this data.

To analyze overall trends in the content severity, we calculate the content severity trend line for each user and then analyze the overall trend line statistics for the users. We observed that many users have steady trend lines with a slope of near zero. To eliminate the noise caused by these neutral trends from our analysis, we filter out the users whose content severity trend lines are essentially flat. These users are either the moderators of the forum or are users that show consistent behavior over time. We then analyze how the content severity of the other users with varying content severity changes over time. Table 3.20 shows the statistics for all the trends lines among all the active users. To eliminate trend lines having a slope near zero, we consider a threshold. We analyze results based on different values of this threshold. For example, for the threshold τ, the corresponding row on the Table only considers trend lines with slope m such that $m < -\tau$ or $m > \tau$ and filters out all other lines having $|m| <= \tau$. We also show the results in the case that there is no threshold (last row of the Table). FLAGGED vs GREEN corresponds to plots with numeric severity value of 1.0 for a FLAGGED post and 0.0 for a GREEN post; Fine-grained severity categories corresponds to plots with following numerics severity values: CRISIS = 1.0, RED = 0.66, AMBER = 0.33, GREEN = 0.0. As illustrated in Table 3.20, we observe an average negative trend line slope for all the values of the threshold. This indicates a decline of average content severity among all the users. Furthermore, we observe that majority of users have a trend line with a negative slope and thus, decreasing severity of content.

These results indicate that overall there is a decline in the content severity of the users as they interact with the forum, which could be due to the potential positive effect of the forum on its users. This effect could be attributed to the users expressing their feelings and emotions, receiving support and feedback from the moderators, and discussing issues with users experiencing similar problems. However, we note that here we only observe the negative trend of content severity; to study the exact causal relationship between interaction with the forum and content severity, further controlled trials on the forum users should be conducted.

Q-2. How the duration of engagement with the forum is associated with users?

We analyze how the duration of a user's engagement with the forum impacts the severity of their posts over time. Tables 3.21a and 3.21b show that users with a first month severity of FLAGGED or URGENT posts interacted with the forum for 3-4 months, while other users interacted with the forum for 5-7 months. These tables are essentially showing that users with less critical posts in the first month tend to interact with the forum in a more long-term basis in comparison with users whose initial posts are critical. The difference in the duration of user interaction by their initial content severity indicates that there are users who visit the forum for immediate assistance in a critical moment and those who use the forum as a longer-term support resource. This result suggests that users whose first posts are more severe could be on the forum for immediate support and will only stay active until their critical mental state reaches a safe equilibrium again. In contrast, the users whose first month is GREEN or non-URGENT may be seeking a long-term resource and a community of users with shared experiences.

This difference between the activity period of users by their initial content reveals an opportunity for moderators to improve their response time to FLAGGED and URGENT posts. Faster moderator attention to FLAGGED and URGENT posts would provide better quality of help to these short-term users and encourage them to further interact with the forum for receiving support. Triaging the forum posts to allow moderators improve their response time would benefit all user groups, and particularly users who currently visit the forum for an immediate support.

Q-3. What is the impact of moderator response time on the user's forum behavior?

Since the focus of this research is on triaging the severity of mental health forum posts, we seek to understand how quickly moderators are currently responding to posts by their severity. Table 3.21 shows the average time for a moderator to respond, as well as the percentage of cases in which the moderators were the first to respond to a user. It shows that in cases where a moderator was the first to respond to a FLAGGED or URGENT post, they took on average more than four hours to respond. Unfortunately, four hours might be too long for users with imminent risks and it is very important to reduce this response time to prevent a potential self harm. Additionally, we observe that moderators are the first responders on less than 33% of non-GREEN posts, meaning the other forum users responded to majority of posts earlier than moderators. This further stresses the value of triaging content severity, so that moderators can quickly respond to critical posts rather than having to identify such posts on the forum manually.

3.4.4 Depression risk assessment in online forums

In previous section, we discussed our approaches for identifying self-harm and suicide signs in mental-health user posts through their language usage. As discussed, these forums are targeted for individuals with mental and emotional issues. Yet, a related but different challenge is to identify depressed users on social media through their posts which are not on any specialized mental-health forums. One example application for this problem would be to provide users experiencing such mental and emotional issues, with the resources they need before some of them reach the extreme case of suicide and self-harm. We argue that it is indeed possible to identify depressed users through their language usage without using any specific mental-health related keywords.

Identifying signs of depression in general social media is a difficult problem that has applications for both better understanding the relationship between mental health and language, and for monitoring a specific user's state (e.g., in the context of monitoring a user's response to clinical care).

To achieve this goal, we propose a general neural network architecture for combining user posts into a representation of a user's activity that is used to classify the user. We further introduce a large-scale novel Reddit dataset that is substantially larger than the existing data and has a much more realistic number of control users. The dataset contains over 9,000 users with self-reported depression diagnoses matched with over 107,000 control users. We show that our approach could be also used to perform self-harm risk assessment on mental health specific forums.

3.4.4.1 Data

Depression dataset construction. We created a new dataset to support the task of identifying forum users with self-reported depression diagnoses. The Reddit Self-reported Depression Diagnosis (RSDD) dataset was created by annotating users from a publicly-available Reddit dataset[15]. Users to annotate were selected by identifying all users who made a post between January 2006 and October 2016 matching a high-precision diagnosis pattern.[16] Users with fewer than 100 posts made before their diagnosis post were discarded. Each of the remaining diagnosis posts was then viewed by three layperson annotators to decide whether the user was claiming to have been diagnosed with depression; the most common false positives included hypotheticals (e.g., "if I was diagnosed with depression"), negations (e.g., "it's not like I've been diagnosed with depression"), and quotes (e.g., "my brother announced 'I was just diagnosed with depression'"). Only users with at least two positive annotations were included in the final group of diagnosed users.

A pool of potential control users was identified by selecting only those users who had *(1)* never posted in a subreddit related to mental health, and *(2)* never used a term related to depression or mental health. These restrictions minimize the likelihood that users with depression are included in the control group. In order to prevent the diagnosed users from being easily identified by the usage of specific keywords that are never used by the control users, we removed all posts by diagnosed users that met either one of the aforementioned conditions (i.e., that was posted in a mental health subreddit or included a depression term).

For each diagnosed user and potential control user, we calculated the probability that the user would post in each subreddit (while ignoring diagnosed users' posts made

[15] `https://files.pushshift.io/reddit/`

[16] e.g., "I was just diagnosed with depression."

to mental health subreddits). Each diagnosed user was then greedily matched with the 12 control users who had the smallest Hellinger distance between the diagnosed user's and the control user's subreddit post probability distributions, excluding control users with 10% more or fewer posts than the diagnosed user. This matching approach ensures that diagnosed users are matched with control users who are interested in similar subreddits and have similar activity levels, preventing biases based on the subreddits users are involved in or based on how active the users are on Reddit. This yielded a dataset containing 9,210 diagnosed users and 107,274 control users. On average each user in the dataset has 969 posts (median 646). The mean post length is 148 tokens (median 74).

The Reddit Self-reported Depression Diagnosis (RSDD) dataset differs from prior work creating self-reported diagnoses datasets in several ways: it is an order of magnitude larger, posts were annotated to confirm that they contained claims of a diagnosis, and a realistic number of control users were matched with each diagnosed user. The lists of terms related to mental health, subreddits related to mental health, high-precision depression diagnosis patterns, and further information are available[17]. We note that this dataset has some (inevitable) caveats: *(i)* the method only captures a subpopulation of depressed people (i.e. those with self-reported diagnosis), *(ii)* Reddit users may not be a representative sample of the population as a whole, and *(iii)* there is no way to verify whether the users with self-reported diagnoses are truthful.

Self-harm assessment. For self-harm risk assessment we use data from mental health forum posts from ReachOut.com, which is a successful Australian support forum for young people. In addition to providing peer-support, ReachOut moderators and trained volunteers monitor and participate in the forum discussions. As discussed

[17] `http://ir.cs.georgetown.edu/data/reddit_depression/`

in previous section, the annotations consist of one of four labels: green (indicating no action is required from ReachOut's moderators), amber (non-urgent attention is required), red (urgent attention is required), and crisis (a risk that requires immediate attention).

Ethical concerns. Social media data are often sensitive, and even more so when the data are related to mental health. Privacy concerns and the risk to the individuals in the data should always be considered [15, 117, 252]. We note that the risks associated with the data used in this work are minimal. This assessment is supported by previous work on the ReachOut dataset [175], on Twitter data [68], and on other Reddit data [160]. The RSDD dataset contains only publicly available Reddit posts. Annotators were shown only anonymized posts and agreed to make no attempts to deanonymize or contact them. The RSDD dataset will only be made available to researchers who agree to follow ethical guidelines, which include requirements not to contact or attempt to deanonymize any of the users. Additionally, for the ReachOut forum data that was explicitly related to mental health, the forum's rules require the users to stay anonymous; moderators actively redact any user identifying information.

3.4.4.2 Methodology

We describe a general neural network architecture for performing text classification over multiple input texts. The model intuitively identifies signals across a user's posts that contribute to their mental health condition; these signals are then merged to derive a vector representation of the user that can be classified into the respective risk category. We propose models based on this architecture for performing two tasks in the social media and mental health domains that we call *self-harm risk classification* and *detecting depression*. The task of self-harm risk classification is estimating a user's

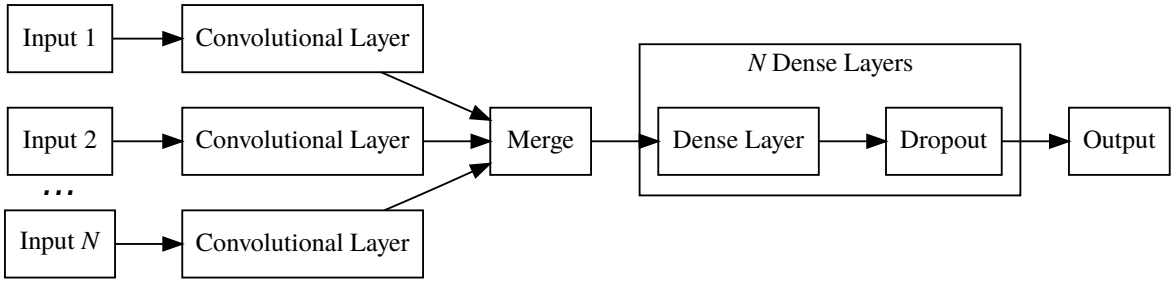


Figure 3.11: Model architecture. The general neural network architecture shared among our user and post classification models. Each input (e.g., each of a user's posts) is processed by a convolutional network and merged to create a vector representation of the user's activity. This vector representation is passed through one or more dense layers followed by an output layer that performs classification. The type of input received, merge operation, and output layer vary with the specific model.

current self-harm risk given the user's post on a mental health support forum and the previous posts in the thread. The task of detecting depression in users is identifying Reddit users with self-reported depression diagnoses given the users' post histories (excluding posts containing mental health keywords or posted in subreddits related to mental health).

While both tasks are focused on predicting a user's mental health status, they differ in both the type of classification performed (i.e., estimating severity on a four point scale vs. boolean classification) and in the amount of data available. Our general architecture serves two purposes: *(1)* identifying relevant features in each input text, and *(2)* combining the features observed in the model's inputs to classify the user.

Shared Architecture. Our proposed models share a common architecture that takes one or more posts as input, processes the posts using a convolutional layer to identify features present in sliding windows of text, merges the features identified into a vector

representation of the user's activity, and uses a series of dense layers to perform classification on the merged vector representation. The type of merging performed and the output layers are properties of the model variant, which we describe in detail in the following section. Convolutional networks have commonly been applied to the task of text classification, such as by Kim [139]. We use categorical cross-entropy as a loss function with both methods, but also experiment with other loss functions when performing severity classification.

First, the model takes one or more posts as input and processes each post with a convolutional network containing a convolutional layer and a pooling layer. The convolutional layer applies filters to a sliding window of k terms *(a)* and outputs a feature value for each sliding window region and each filter *(b)*. The same filters are applied to each window; each filter can be viewed as a feature detector and the overall process can be conceptualized as looking for windows of terms that contain specific features. The features are not specified a priori through feature engineering, but instead are learned automatically when the model is trained. After identifying the features present in each region (i.e., sliding window), a max pooling layer considers non-overlapping regions of length n and keeps the highest feature value for each region *(c)*. This step eliminates the regions (i.e., sliding windows) that do not contain useful features, which reduces the size of the convolutional network's output. The same convolutional network is applied to each input post, meaning that the model learns to look for the same set of features in each.

After each input post has been processed by a convolutional network, the output of each convolutional network is merged to create a representation of the user's activity across all input posts. This representation is processed by one or more dense layers (i.e., fully connected layers) with dropout [243] before being processed by a final output layer to perform classification. The type of output layer is dependent on the

Table 3.23: The hyperparameters used by each model.

Method		Convolution			Dense Layers	Dropout	Class Balance
		Size	Filters	Pool Len.			
Reddit	Cat. Cross Ent.	3	25	all (avg)	1 w/ 50 dims	0.0	Sampled
ReachOut	Cat. Cross Ent.	3	150	3 (max)	2 w/ 250 dims	0.3	Weighted
	MSE	3	100	3 (max)	2 w/ 250 dims	0.5	Sampled
	Class Metric	3	100	3 (max)	2 w/ 150 dims	0.3	Sampled

model variant. Our shared model architecture is illustrated in Figure 3.11. The architecture's hyperparameters (e.g., the sliding window size k, the number of convolutional filters used, and type of pooling) also vary among models and are described in §3.4.4.4. Both the convolutional and dense layers use ReLU activations [183] in all model variants.

3.4.4.3 Models

Depression detection. Our model for depression detection takes a user's posts as input and processes each post with a convolutional network. Each convolutional network performs average pooling to produce its output. These post representations are then merged with a second convolutional layer to create a user representation; we found this approach led to more stable performance than using a second average pooling or max pooling layer. The user representation created by the merge step is then passed to one or more dense layers before being passed to a dense output layer with a softmax activation function to perform classification. The number of dense layers used is a hyperparameter described in §3.4.4.4. Categorical cross-entropy is used as the model's loss function.

While our model shares similarities with CNN-based models in prior work [134, 139, 270], it focuses on learning representations of user's posts and combining the post representations into an overall representation of the user's activity.

Self-harm risk assessment. Our model for self-harm risk classification takes two inputs: the target post being classified and the prior posts (if any) in the target post's thread. The prior posts provide context and are thus useful for estimating the risk of self-harm present in the target post. The two inputs are both processed by a convolutional network as in user-level classification, but in this case the convolutional network's outputs correspond to a representation of the target post and to a representation of the target post's context (i.e., the prior posts in the thread). Given that these two outputs represent different aspects, they are merged by concatenating them together. This merged representation is then passed to one or more dense layers and to an output layer; the type of output layer depends on the loss function used. There are four self-harm risk assessment model variants in total:

Categorical Cross Ent. uses an output layer with a softmax activation function, and categorical cross-entropy as its loss function. This mirrors the output layer and loss function used in the user level classification model.

MSE uses an output layer with a linear activation function, and mean squared error as its loss function. The model's output is thus a single value; to perform classification, this output value is rounded to the nearest integer in the interval $[0, t-1]$, where t is the number of target classes.

The final two loss functions perform metric learning rather than performing classification directly. They learn representations of a user's activity and of the four self-harm risk severity labels; classification is performed by comparing the euclidean

distance between a representation of a user's activity (produced by the final layer) and each of the four severity label representations.

Class Metric: Let d be the size of the output layer and X be the layer's d-dimensional output. *Class Metric* learns a d-dimensional representation of each class C_i such that $||X - C_i||_2$ is minimized for the correct class i; this is accomplished with the loss function:

$$L_{i,p,n} = \max\left(0, ||X_i - C_p||_2 - ||X_i - C_n||_2 + \alpha\right)$$

where C_p is the correct (i.e., positive) class for X_i, C_n is a randomly chosen incorrect (i.e., negative) class, and α is a constant to enforce a minimum margin between classes. Classification is performed by computing the similarity between X_i and each class C_j.

Class Metric (Ordinal) extends *Class Metric* to enforce a margin between ordinal classes as a function of the distance between classes. Given a ranked list of classes such that more similar classes have closer rankings, that is $\forall i$ $sim(C_i, C_{i\pm1}) > sim(C_i, C_{i\pm2})$, we incorporate the class distance into the margin such that more distant incorrect class labels must be further away from the correct class label in the metric space. The loss function becomes

$$L_{i,p,n} = \max\left(0, \ \ ||X_i - C_p||_2 \ \ - ||X_i - C_n||_2 + \alpha|p - n|\right)$$

where $|p - n|$ causes the margin to scale with the distance between classes p and n.

3.4.4.4 Experiments

In this section, we describe the model hyperparameters used and present our results on the depression detection and self-harm risk assessment tasks. To facilitate reproducibility we provide our code and will provide the Reddit depression dataset to researchers who sign a data usage agreement[18].

[18] `http://ir.cs.georgetown.edu/data/reddit_depression/`

Experimental setup. The hyperparameters used with our models are shown in Table 3.23. The severity risk assessment models' hyperparameters were chosen using 10-fold cross validation on the 947 ReachOut training posts, with 15% of each fold used as validation data. The depression identification model's hyperparameters were chosen using the Reddit validation set. The depression identification model's second convolutional layer (i.e., the layer used to merge post representations) used filters of length 15, a stride of length 15, and the same number of filters as the first convolutional layer. All models were trained using stochastic gradient descent with the Adam optimizer [142]. The hyperparameters that varied across models are shown in Table 3.23. The convolution size, number of convolutional filters, pooling type, pooling length, and number of dense layers was similar across all post models. Class balancing was performed with *Categorical Cross Ent.* by weighting classes inversely proportional to their frequencies, whereas sampling an equal number of instances for each class worked best with the other methods.

Addressing limited data. The post classification models' input consists of skip-thought vectors [143]; each vector used is a 7200-dimensional representation of a sentence. Thus, the convolutional windows used for post classification are over sentences rather than over terms. This input representation was chosen to mitigate the effects of the ReachOut dataset's relatively small size. The skip-thought vectors were generated from the the ReachOut forum dataset by sequentially splitting the posts in the training set into sentences, tokenizing them, and training skip-thoughts using Kiros et al.'s implementation with the default parameters. Sentence boundary detection was performed using the Punkt sentence tokenizer [145] available in NLTK [20]. These 2400-dimensional forum post skip-thought vectors were concatenated with the 4800-dimensional book corpus skip-thought vectors available from Kiros et al.. Exper-

iments on the training set indicated that using only the ReachOut skip-thought vectors slightly decreased performance, while using only the book corpus skip-thought vectors substantially decreased performance. As input the post models received the last 20 sentences in each target post and the last 20 sentences in the thread prior to the target post; any prior sentences are ignored.

Depression detection. The data used for depression detection was described in §3.4.4.1. As baselines we compare our model against the FastText classifier [131] and MNB and SVM classifiers [266] using features from prior work. We tune FastText's hyperparameters on the validation set. Specifically, we consider a maximum n-gram size $\in [1, 2, 3, 4, 5]$, an embedding size $\in [50, 100, 150]$, and a learning rate $\in [0.05, 0.1, 0.25, 0.5]$ as suggested in the documentation. We consider two sets of features for the MNB and SVM classifiers. The first set of features is the post content itself represented as sparse bag of words features (*BoW baselines*). The second set of features (*feature-rich baselines*) comprises a large set of features including bag of words features encoded as sparse weighted vectors, external psycholinguistic features captured by LIWC[19] [208], and emotion lexicon features [245]. Since our problem is identifying depression among users, psycholinguistic signals and emotional attributes in the text are potentially important features for the task. These features have been also previously used by successful methods in the Twitter self-reported diagnosis detection task [68]. Thus, we argue that these are strong baselines for our self-reported diagnosis detection task. We apply count based and TF-IDF based feature weighting

[19] `http://liwc.wpengine.com/`

for bag of words features. We perform standard preprocessing by removing stopwords and lowercasing the input text.[20]

The data is split into training, validation, and testing datasets each containing approximately 3,000 diagnosed users and their matched control users. The validation set is used for tuning development and hyperparameter tuning of our models and the baselines. The reported results are on the test set. The depression detection models' input consisted of raw terms encoded as one-hot vectors. We used an input layer to learn 50-dimensional representation of the terms. For each target user, the CNN received up to n_{post} posts containing up to n_{term} terms. In this section we present results for two values of n_{post}. The earliest post approach (CNN-E) takes each user's $n_{post} = 400$ earliest posts as input. The random approach (CNN-R) samples $n_{post} = 1500$ random posts from each user. We empirically set $n_{term} = 100$ with both approaches. We later analyze the model's performance as n_{post} and n_{term} vary in §3.4.4.5 and as the post selection strategy varies in §3.4.4.5.

Results. The results of identifying depressed users for our model and baselines are shown in Table 3.24. Our proposed model outperforms the baselines by a large margin in terms of recall and F1 on the diagnosed users (increases of 41% and 16%, respectively), but performs worse in terms of precision. As described later in the analysis section, the CNN identifies language associated with negative sentiment across a user's posts.

Self-harm risk classification. We also show the effectiveness of our model on the task of self-harm risk assessment. We train our methods to label the ReachOut posts

[20]During experimentation, we found TF-IDF sparse feature weighting to be superior than other weighting schemes. Additional features such as LDA topics and χ^2 feature selection did not result in any further improvements.

Table 3.24: Performance of identifying depressed users on the Reddit test set.

Method	Precision	Recall	F1
BoW - MNB	0.44	0.31	0.36
BoW - SVM	0.72	0.29	0.42
Feature-rich - MNB	0.69	0.32	0.44
Feature-rich - SVM	0.71	0.31	0.44
FastText	0.37	**0.70**	0.49
User model - CNN-E	0.59	0.45	0.51
User model - CNN-R	**0.75**	0.57	**0.65**

The differences between the CNN and baselines are statistically significant (McNemar's test, $p < 0.05$).

and compare them against the top methods from CLPsych '16. We use the same experimental protocol as was used in CLPsych '16; our methods were trained on the 947 training posts and evaluated on the remaining 280 testing posts. We used 15% of the 947 training posts as validation data.

We report results using the same metrics used in CLPsych, which were: the macro-averaged F1 for the *amber*, *red*, and *crisis* labels (*non-green* posts); the macro-averaged F1 of *green* posts vs. *amber* $\cup$ *red* $\cup$ *crisis* (*flagged* posts); and the macro-averaged F1 of *green* $\cup$ *amber* vs. *red* $\cup$ *crisis* (*urgent* posts). The *non-green* F1 was used as the official CLPsych metric with the intention of placing emphasis on classification performance for the non-green categories (i.e., those that required some response). The binary *flagged* meta-class was chosen to measure models' abilities to differentiate between posts that require attention and posts that do not, and the

Table 3.25: Self-harm risk assessment performance on the ReachOut CLPsych test set.

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Baseline [175]	0.31	0.78	0.86	0.38	0.89	-	-
Kim et al. [138]	0.42	0.85	0.91	0.62	0.91	0.55	0.85
Malmasi et al. [167]	0.42	0.87	0.91	0.64	0.93	0.55	0.83
Brew [29]	0.42	0.78	0.85	0.69	0.93	0.54	0.79
Cohan et al. [54]	0.41	0.81	0.87	0.67	0.92	0.53	0.80
Categorical Cross Ent.	**0.50**	**0.89**	**0.93**	0.70	**0.94**	**0.61**	**0.89**
MSE	0.42	0.80	0.85	0.64	0.93	0.53	0.78
Class Metric	0.46	0.79	0.84	0.70	**0.94**	0.56	0.80
Class Metric (Ordinal)	0.47	0.88	**0.93**	**0.72**	0.93	0.59	0.87

Results for the other methods are from [175]. Differences in performance between the following pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *Class Metric*, *MSE* and *Categorical Cross Ent.*, *MSE* and *Class Metric (Ordinal)*, and *Class Metric (Ordinal)* and *Class Metric*.

binary *urgent* meta-class was chosen to measure their abilities to differentiate between posts that require quick responses and posts that do not. In addition to macro-averaged F1, CLPsych also reported the accuracy for each category. We additionally report F1 macro-averaged over all classes.

Results. The results on the self-harm risk assessment task for our models and for the current best-performing methods (briefly explained in §3.4.2) are shown in Table 3.25. We also report a baseline result which is based on a SVM classifier with bigram

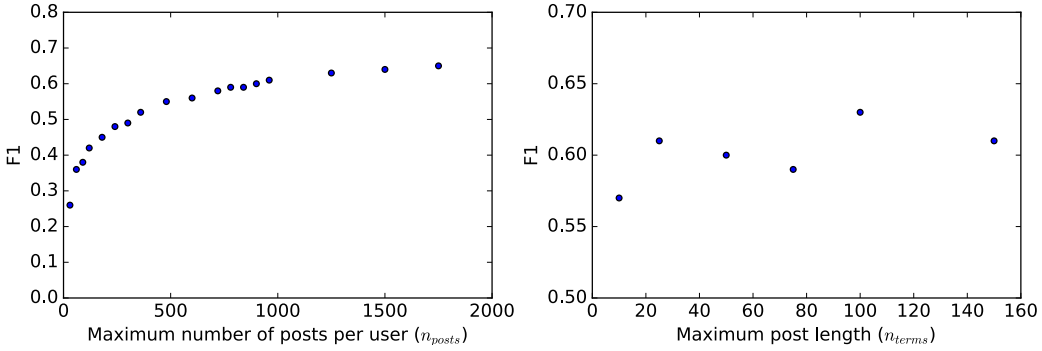
Table 3.26: Self-harm risk assessment performance on CLPsych training set (10-told cross validation).

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Categorical Cross Ent.	0.54	0.87	0.89	0.69	0.91	0.63	0.80
MSE	**0.87**	**0.95**	**0.96**	**0.91**	**0.98**	**0.89**	**0.93**
Class Metric	0.73	0.90	0.91	0.81	0.94	0.78	0.86
Class Metric (Ordinal)	0.85	**0.95**	**0.96**	0.89	0.97	0.88	0.92

Categorical Cross Ent. performs substantially worse than on the test set, while *MSE* performs substantially better. *Class Metric (Ordinal)* continues to perform well. The difference in performance between the following method pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *MSE*, *Categorical Cross Ent.* and *Class Metric*, *Categorical Cross Ent.* and *Class Metric (Ordinal)*, *MSE* and *Class Metric*, and *Class Metric* and *Class Metric (Ordinal)*.

features. When measured by *non-green* F1, the official metric of the CLPsych '16 Triage Task, our proposed models perform up to 19% better than the best existing methods. Similarly, our models perform up to 11% better when measured with an F1 macro-averaged across all categories (i.e., *all* column) and up to 5% better with measured accuracy across all categories. *Categorical Cross Ent.* performs best in all of these cases, though the difference between the performance of *Categorical Cross Ent.* and *Class Metric* with an ordinal margin is not statistically significant.

We also evaluate the performance of our methods on the training set using 10-fold cross validation to better observe performance differences (Table 3.26). All model variants perform substantially better on the training set than on the test set. This is partially explained by the fact that the models were tuned on the training set, but the



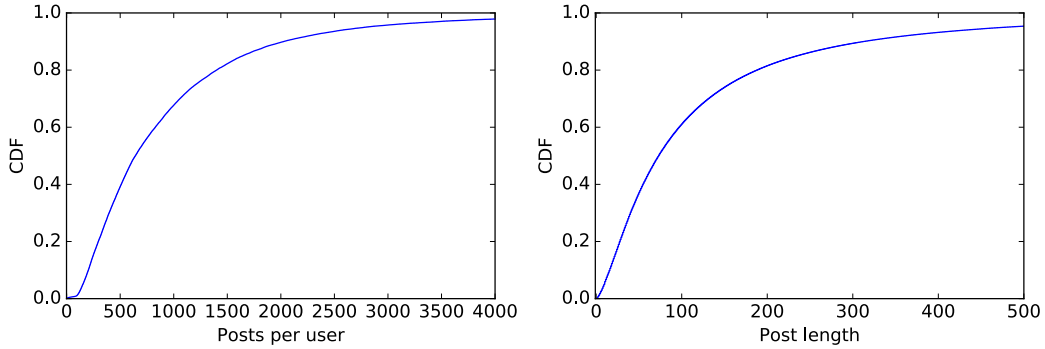
(a) **(b)**

Figure 3.12: Sensitivity of the CNN-R model to the parameters. Sensitivity of the CNN-R model to the parameters n_{posts} (a) and n_{terms} (b) on RSDD's validation set. F1 increases as n_{posts} does (a), but the rate of increase slows as n_{posts} surpasses 1000. The trend for n_{terms} is less clear (b), but the highest F1 is achieved at $n_{terms} = 100$. In Figure (a) the parameter n_{terms} was fixed to 100, and in Figure (b) n_{posts} was fixed to 1500.

large difference in some cases (e.g., the increase in the highest non-green F1 from 0.50 to 0.87) suggest there may be qualitative differences between the datasets. The best-performing method on the test set, *Categorical Cross Ent.*, performs the worst on the training set; worst-performing method on the test set, *MSE*, performs the best on the training set. *Class Metric (Ordinal)* performs well on both the testing and training sets, however, suggesting that it is more robust than the other methods. Furthermore, there is no statistically significant difference between *Class Metric (Ordinal)* and the best-performing method on either dataset.

3.4.4.5 Analysis

Posts per user and post length. In this section we consider the effects of the maximum number of posts per user (i.e., n_{post}) and the maximum post length (i.e., n_{term})



(a) **(b)**

Figure 3.13: Empirical cumulative distribution functions (CDF) of the number of posts per user (a) and the post length (b) in the RSDD dataset.

on the Reddit dataset. To do so we train the CNN-R model as described in §3.4.4.4 and report F1 on the validation set. When varying n_{post} we set $n_{term} = 100$, and when varying n_{term} we set $n_{post} = 1500$.

As shown in Figure 3.12, the best performance of the CNN-R model is reached when it considers 100 terms in posts and up to 1750 posts for each user. F1 increases as n_{post} increases, up to the maximum tested value of 1750 (Figure 3.12a). There is relatively little change in F1 from $n_{post} = 1250$ to $n_{post} = 1750$, however, so we use $n_{post} = 1500$ in our experiments for efficiency reasons. As shown in Figure 3.13a, approximately 20% of users have more than 1500 posts. The effect of the maximum post length is not consistent (Figure 3.12b), but performance is maximized at $n_{term} = 100$. As shown in Figure 3.13b, approximately 40% of posts are longer than 100 terms.

Post selection. For users with more than the maximum number of posts n_{post}, a post selection strategy dictates which posts are used as input to the model. Table 3.27 shows the effect of the post selection strategy on the Reddit dataset's validation set.

Table 3.27: Models' performance on RSDD's validation set with different post selection strategies and values of n_{post}.

Post Selection	$n_{posts} = 400$			$n_{posts} = 1500$		
	Precision	Recall	F1	Precision	Recall	F1
Earliest	**0.58**	0.46	0.52	0.59	0.55	0.57
Latest	**0.58**	0.50	**0.54**	0.69	**0.59**	0.64
Random	0.52	**0.53**	0.53	**0.71**	**0.59**	**0.65**

CNN-E corresponds to the earliest strategy with $n_{post} = 400$ and CNN-R corresponds to the random strategy with $n_{post} = 1500$.

Selecting a user's earliest posts performs the worst regardless of n_{post}'s value, though the differences in F1 are smaller when $n_{post} = 400$. Randomly selecting posts for each user performs the best across all metrics when $n_{post} = 1500$, with a large increase in precision over selecting users' earliest posts and a small increase over choosing users' latest posts.

Phrases contributing to classification. In this section we analyze the language that strongly contributed to the identification of depressed users on the Reddit dataset. Unfortunately, it is impossible to show entire Reddit posts without compromising users' anonymity; we found that even when a post is paraphrased, enough information remains that it can easily be identified using a Web search engine. For example, one Reddit post that strongly contributed to the author's classification as a depressed user contained the mention of a specific type of abuse and several comments vaguely related to this type of abuse. We attempted to paraphrase this post, but found that

Table 3.28: Example phrases that strongly contributed to user classification.

Top Phrases	
i went to	to scare you
my whole	to have it
sometimes i	my son was
i'm so sorry	it wasn't

any paraphrase containing general language related to both the type of abuse and to the user's comments was enough to identify the user. Thus, to protect the anonymity of the users in our dataset, we do not publish posts in any form.

Rather than publishing posts, we identify key phrases in posts from users who were correctly identified as being depressed. Phrases from eight self-reported depressed users are shown in Table 3.28; to prevent these phrases from being used to identify users, we retain only the top phrase from each user. These phrases were identified by using the model's convolutional filter weights to identify posts in the validation dataset that are strongly contributing to the model's classification decision, and then using the convolutional filter weights to identify the phrase within each post that most strongly contributed to the post's classification (i.e., had the highest feature values).

In keeping with the design of our dataset, terms related to depression or diagnoses are not present. Instead, the model identifies phrases that often could be associated with a negative sentiment or outlook. For example, "my whole" could be part of a negative comment referring to the poster's whole life. It should be noted that the

Table 3.29: Self-harm risk assessment performance on CLPsych '17 test set.

Method	Non-green	Flagged		Urgent		All	
	F1	F1	Acc.	F1	Acc.	F1	Acc.
Categorical Cross Ent.	**0.37**	0.88	**0.90**	0.48	0.83	**0.50**	**0.71**
MSE	0.31	0.84	0.84	**0.54**	**0.84**	0.44	0.64
Class Metric	0.30	0.88	0.89	0.46	0.81	0.45	0.68
Class Metric (Ordinal)	0.34	**0.89**	**0.90**	0.49	0.81	0.48	0.69

All methods perform substantially worse than on the CLPsych '16 test data. The difference in performance between the following method pairs are statistically significant (McNemar's test, $p < 0.05$): *Categorical Cross Ent.* and *MSE*, and *MSE* and *Class Metric (Ordinal)*.

model makes classification decisions based on the occurrence of phrases across many posts by the same user. Though one can imagine how the phrases shown here could be used to convey negative sentiment, the presence of a single such phrase is not sufficient to cause the model to classify a user as depressed.

CLPsych '17 shared task. In this section we report results on the 2017 CLPsych Workshop's self-harm risk classification task.[21] While CLPsych '17 featured the same self-harm risk classification task as CLPsych '16 (§3.4.4.4), new test data was used to conduct the evaluation. This provides an opportunity to further evaluate our model on

[21]The 2017 test data was released after the initial version of this manuscript had been completed. An official overview paper for CLPysch '17 is not yet available at the time of writing.

the task of self-harm risk assessment and to conduct an error analysis. The methods were configured and evaluated in the same manner as described in §3.4.4.4.[22]

Results are shown in Table 3.29. All methods perform substantially worse than they performed on the CLPsych '16 test data as measured by non-green, urgent, and overall F1. The trends across methods remain similar, however, with *Categorical Cross Ent.* performing the best as measured by non-green and overall F1, and with no statistically significant difference between *Class Metric (Ordinal)* and the best performing method.

Notably, the methods' flagged F1 scores do not see a similar decrease on the CLPsych '17 data. This suggests that the decreased performance is being caused by an inability to distinguish between the non-green classes (i.e., amber, red, and crisis). The importance of differentiating between the red and crisis classes increased with the 2017 shared task, because the proportion of crisis labels in the data increased from 0.4% (2016 testing) and 4% (2016 training) to 11% (2017 testing). The methods rarely classify a post as crisis, however, causing an increase in the number of misclassifications on the 2017 testing data. For example, *Class Metric (Ordinal)* classified only four posts from the 2017 test data as crisis, and it classified no posts from the 2016 test data as crisis. We leave improving the model to better identify crisis posts as future work.

3.5 Conclusions

The growing amount of clinical data and electronic health records in medical centers requires automated processing for purposes such as improving health care, public

[22]The results in this section differ slightly from the methods' results as reported by CLPsych '17. Here the methods were trained on only CLPsych '16 training data to match the experimental setup described earlier, whereas the methods were trained on both the CLPsych '16 training and test data in the official results reported by CLPsych '17.

health surveillance, and improving medical education. This chapter first described our proposed methods for identifying significant discrepancies in clinical reports which is essential for patient care and resident education. The proposed method is a two-stage pipeline to distinguish between significant and non-significant discrepancies in radiology reports. The first stage adopted a heuristic approach based on the RadLex domain ontology and negations in radiology narratives. The second stage is a classifier based on several features including summarization and machine translation evaluation, and text readability features for classification of the reports. This method was validated using a real-world dataset of medical radiology reports obtained from a large urban hospital. On this dataset, I showed the effectiveness of our method with statistically significant improvement (+14.6% AUC, -52% FNR) over several baselines. A patent application based on the proposed approach has been filed at United States Patent and Trademark Office[23].

I then presented a neural network model for identifying harm in clinical narratives related to healthcare. This method consists of a multi-layer neural network with convolutional, recurrent, and soft attention mechanism layers. I argued that the convolutional layer is important in finding the local features and the recurrent layer with attention is effective in finding the interactions and dependencies along the sequence. I demonstrated that this method can significantly improve the performance over existing methods in identifying harm safety cases. The impact of the methods and results presented in this work is substantial to patient care. More accurate methods in the identification of harm can help the data analysis and reporting process, prevent harm to patients, better prioritize resources to address safety incidents, and subsequently improve general patient care.

[23]https://patents.google.com/patent/US20170206317A1/en

In final section of this chapter, I argued for the close connection between social media and mental health, and described how we can use NLP and text categorization methods to attempt to address important challenges in mental health. I presented an approach for assessing the user content severity in mental peer support forums with a specific goal of identifying cases with potential risk of self-harm or suicide. To achieve this goal, I used a feature-rich classifier with various sets of features including psycholinguistic, contextual, topic modeling and forum metadata features for triaging the content into different severity categories. In addition to a single classifier, I also built an ensemble classifier by using different sets of features. I evaluated this approach on real-world data from a large mental health forum, ReachOut.com. Our method effectively improves over the state-of-the-art by large margins (up to 17% macro-average F1 scores of critical categories). I showed that the content severity of the users tend to decrease as they interact with the forum. Results further indicated that there is a need for effective and efficient triaging of forum post data to assist the moderators in attending the users with potential risk of self-harm. This research stresses the importance of mental health forums as a support platform for users with mental health problems. It furthermore provides an efficient and effective way for moderators to asses the content severity of the forum, and consequently help individuals in need and prevent self harm incidents.

I further described a CNN based neural network architecture for identifying depression through general language usage in general online forums. I also described the construction of the Reddit Self-reported Depression Diagnosis (RSDD) dataset, a large-scale dataset, containing over 9,000 users with self-reported depression diagnoses matched with over 107,000 similar control users. Our dataset is available to the community for further research in this area[24]. Our classification approach on

[24] `http://ir.cs.georgetown.edu/data/reddit_depression/`

the RSDD dataset, substantially outperformed strong existing methods in terms of Recall and F1. While these depression detection results are encouraging, the absolute values of the metrics illustrate that this is a challenging task and worthy of further exploration.

The presented research efforts and outcomes have the following impacts: they provide a strong approach to identifying posts indicating a risk of self-harm in social media; they demonstrate a means for large scale public mental health studies surrounding the state of depression; and they demonstrate the possibility of sensitive applications in the context of clinical care, where clinicians could be notified if the activities of their patients suggest they are at risk of self-harm. Furthermore, large-scale datasets such as our RSDD dataset can provide complementary information to existing data on mental health which are generally relatively smaller collections.

Chapter 4

Conclusions

In this dissertation, I introduced solutions to some of real-world challenges and problems in the health domain and I demonstrated the life-saving potential of NLP in healthcare and medical research. The results of my dissertation can be or are already being used to help doctors, patients and scientists by:

- summarizing scientific articles
- analyzing textual reports of medical errors
- improving the eduction of medical residents
- identifying at-risk individuals in social media

In Chapter 2, I focused on scientific publications as the main source of knowledge dissemination in science and validated my first hypothesis (*H1*) on improving scientific document summarization. The publication rate of scientific papers has been constantly increasing in past decades, making it challenging for researchers to keep up with the new developments. One solution is to utilize automatic text summarization methods to summarize key contributions and findings of scientific papers. Previous work shows how a set of citation texts describing a referenced paper can be leveraged to provide a summary of the main contributions of the paper. I argued how contributions of a paper distilled into a summary via citations can be inaccurate in conveying

the exact content of the original paper. This is due to citations sometimes inaccurately ascribing contributions to the referenced paper or quoting results without mentioning the assumptions or conditions. The problem of "lacking reference context" is even more important in health sciences as findings of papers can have impact on human lives. To address these challenges, I presented effective contextualization methods for citation texts by linking them to their relevant textual parts in the referenced article. The proposed methods of citation contextualization were threefold: a query reformulation based method, an information retrieval model based on word embeddings and domain knowledge, and supervised classification. All these approaches were found to be substantially better performing than other methods in contextualizing citations. I then showed how these contextualized citations along with the discourse structure of the scientific document can be leveraged to generate the summary of contributions of the paper. This type of summary addresses the shortcoming of existing citation-based summarization methods. On two scientific datasets from both health sciences and computational linguistics domain, my proposed methods signficantly outperformed the existing methods. My citation-based summarization approach was extractive, meaning it generates the summary by selecting sentences from the input. In addition to this method, I presented an abstractive summarization approach that generates the summary from scratch and without necessarily copying words or phrases from the input. This approach was based on a discourse-aware sequence-to-sequence attention model that uses scientific document discourse structure for generating the summary. On two large-scale datasets of scientific papers I showed how this method effectively outperforms the state-of-the-art abstractive models in terms of standard summarization evaluation metrics. I concluded chapter 2 by providing an analysis of the effectiveness of current summarization evaluation metrics in the scientific domain

and showed how we can exploit semantic relatedness instead of superficial features to improve summarization evaluation.

In Chapter 3, I focused on other significant sources of textual data in the health domain, specifically, medical reports and social media. This chapter particularly addressed my second major hypothesis (*H2*) on text categorization application in the health-related domain. Medical errors are known to be among leading causes of death world-wide. I presented text classification methods that identify problems leading to medical errors. I first showed how we can differentiate errors from stylistic variations between versions of medical narratives by using carefully designed features such as machine translation and summarization evaluation metrics. On a real-world dataset, this approach resulted in significant improvements of about 15% incearse in Area Under the Curve and 52% decrease in false negative rate over the baselines. Healthcare systems often use reporting systems to track errors and harm to patients. These reports are natural language narratives, describing events that happen to the patient while they are at the healthcare center. While on a case-by-case basis, these reports can shed light on the events that contributed to patient harm, large scale manual analysis of these reports for finding common reasons of harm is impractical. I showed how an attention-based neural classification model can be used to identify harm in the medical reports. By focusing on harm cases, clinical professionals can identify the common reasons for harm much more easily. This data-driven method effectively obtained significantly improved classification scores over existing baselines. As another prominent dimension of healthcare, I studied mental-health through social media to see how Natural Language Processing (NLP) can be used to address some of the problems in mental-health. I showed how in specialized mental-health forums, we can use the language of users to identify posts that manifest signs of suicidal thoughts. By identifying such posts, we can direct the attention of the forum mod-

erators to engage with the user through discussions and provide them with the help they need. In particular, my method compared with the existing systems was able to effectively increase macro-F1 scores of identifying critical categories by 17%. In a more broad setting, instead of identifying cases of suicide or self-harm, my next research goal was to identify depressed users. I first discussed creation of a dataset to support the task of identifying depression through language without specific mentions of depression-related keywords. This resulted in a large-scale dataset of more than 110,000 users and millions of posts that help researchers study the problems of mental-health through social media. I also proposed a classification architecture based on convolutional neural networks to only use the general language of the users to identify depression. I demonstrated how this method achieves significant improvements over strong baselines.

This dissertation provided a summary of my previously published results [48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 221, 273]. My research goal throughout this dissertation was to explore significant challenges that exist in understanding textual health-related data.

To summarize, the impacts of the research questions, datasets, solutions and approaches explored in this thesis are significant from multiple perspectives: *(i)* they provide better summarization methods that will help researchers learn about new findings in the scientific community in a significantly reduced amount of time; *(ii)* they raise the problem of inadequacy of current standard evaluation metrics (i.e., ROUGE framework) for summarization and present an alternative method that can gain higher correlation with manual judgements; *(iii)* they show improved methods of identifying harm in patient-care from clinical narratives that can help the data analysis and reporting processes, prevent harm to patients, better prioritize resources to address safety incidents, and subsequently improve general patient care; *(iv)* they

stress the importance of mental health forums as a support platform for users with mental health problems and provides an efficient and effective method for moderators to asses the content severity of the forum, and consequently help individuals in need and prevent self harm incidents. *(v)* they provide approaches for identifying posts indicating a risk of self-harm in general social media such as reddit. *(vi)* they demonstrate a means for large scale public mental health studies surrounding the state of depression; and they demonstrate the possibility of sensitive applications in the context of clinical care, where clinicians could be notified if the activities of their patients suggest they are at risk of self-harm. *(vii)* large-scale datasets such as the one presented in this thesis, can provide complementary information to existing data on mental-health which are generally relatively smaller collections.

While these impacts are significant, many of the challenges discussed in this dissertation are non-trivial and I only took a few steps towards addressing those. My hope is that my dissertation can attract attention of other interested NLP researchers to further study these real-world problems.

Bibliography

[1] Asad Abdi, Norisma Idris, Rasim M Alguliyev, and Ramiz M Aliguliyev. Pdlk: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications*, 42(22):8936–8946, 2015.

[2] Amjad Abu-Jbara and Dragomir Radev. Coherent citation-based summarization of scientific papers. In *ACL '11*, pages 500–509. Association for Computational Linguistics, 2011.

[3] Amjad Abu-Jbara and Dragomir Radev. Reference scope identification in citing sentences. In *NAACL-HLT*, pages 80–90. ACL, 2012.

[4] Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R Radev. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *NAACL-HLT*, pages 596–606, 2013.

[5] Agency for Healthcare Research and Quality. Categories of medication error classification. *Content last reviewed August*, 2012. URL `http://www.ahrq.gov/`.

[6] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In *Mining text data*, pages 163–222. Springer, 2012.

[7] Tim Althoff, Kevin Clark, and Jure Leskovec. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *arXiv preprint arXiv:1605.04462*, 2016.

[8] American Foundation for Suicide Prevention. Suicide statistics 2016. *AFSP; New York, NY*, 2016.

[9] Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. Success with style: Using writing style to predict the success of novels. *Poetry*, 580(9):70, 2013.

[10] Iana Atanassova, Marc Bertin, Vincent Larivière, and David Bawden. On the composition of scientific abstracts. *Journal of Documentation*, 72(4), 2016.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[12] Michael Bendersky and W Bruce Croft. Discovering key concepts in verbose queries. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 491–498. ACM, 2008.

[13] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[15] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. *EACL 2017*, page 94, 2017.

[16] Adrian Benton, Margaret Mitchell, and Dirk Hovy. Multitask learning for mental health conditions with limited social media data. In *Proceedings of*

the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 152–162, Valencia, Spain, April 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E17-1015.

[17] Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 481–490. Association for Computational Linguistics, 2011.

[18] Marc Bertin, Iana Atanassova, Yves Gingras, and Vincent Larivière. The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, 67(1):164–177, jan 2016. ISSN 23301635. doi: 10.1002/asi.23367. URL http://doi.wiley.com/10.1002/asi.23367.

[19] Akshay Bhat, George Shih, and Ramin Zabih. Automatic selection of radiological protocols using machine learning. In *Proceedings of the 2011 workshop on Data mining for medicine and healthcare*, pages 52–55. ACM, 2011.

[20] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[21] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, 2012.

[22] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[23] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[24] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[25] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11): 2215–2222, apr 2015. doi: 10.1002/asi.23329. URL `http://dx.doi.org/10.1002/asi.23329`.

[26] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. Validating machine learning algorithms for twitter data against established measures of suicidality. *JMIR mental health*, 3(2):e21, 2016.

[27] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 20(2):172–188, 2008.

[28] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[29] Chris Brew. Classifying reachout posts with a radial basis function svm. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 138–142, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[30] Pete Burnap, Walter Colombo, and Jonathan Scourfield. Machine classification and analysis of suicide-related communication on Twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 75–84. ACM, 2015.

[31] Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250. ACM, 2008.

[32] Ziqiang Cao, Wenjie Li, and Dapeng Wu. Polyu at cl-scisumm 2016. In *BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries*, 2016.

[33] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336. ACM, 1998.

[34] Asli Celikyilmaz and Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In *ACL*, pages 815–824. Association for Computational Linguistics, 2010.

[35] Centers for Disease Control and Prevention. Suicide facts at a glance 2015. *CDC; Atlanta, GA: Department of Health and Human Services*, 2015. URL `http://www.cdc.gov/violenceprevention/suicide/statistics/`.

[36] Tanmoy Chakraborty and Ramasuri Narayanam. All fingers are not equal: Intensity of references in scientific articles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1358, Austin, Texas, November 2016. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1142`.

[37] Tanmoy Chakraborty, Amrith Krishna, Mayank Singh, Niloy Ganguly, Pawan Goyal, and Animesh Mukherjee. Ferosa: A faceted recommendation system for scientific articles. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 528–541. Springer, 2016.

[38] Yllias Chali and Sadid a. Hasan. Query-focused multi-document summarization: Automatic data annotations and supervised learning approaches. *Nat. Lang. Eng.*, 18(1):109–145, January 2012. ISSN 1351-3249. doi: 10.1017/S1351324911000167. URL `http://dx.doi.org/10.1017/S1351324911000167`.

[39] Wendy W Chapman, Dieter Hilert, Sumithra Velupillai, Maria Kvist, Maria Skeppstedt, Brian E Chapman, Michael Conway, Melissa Tharp, Danielle L Mowery, and Louise Deleger. Extending the negex lexicon for multiple languages. *Studies in health technology and informatics*, 192:677, 2013.

[40] Wendy Webber Chapman, Gregory F Cooper, Paul Hanbury, Brian E Chapman, Lee H Harrison, and Michael M Wagner. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. *Journal of the American Medical Informatics Association*, 10(5):494–503, 2003.

[41] Liangzhe Chen, K. S. M. Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B. Aditya Prakash. Syndromic surveillance of flu on twitter using weakly supervised temporal topic models. *Data Mining and Knowledge Discovery*, 30(3):681–710, 2016. ISSN 1573-756X. doi: 10.1007/s10618-015-0434-x.

[42] Tianqui Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, 2016. ACM.

[43] Colin Cherry, Saif M Mohammad, and Berry De Bruijn. Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5(Suppl 1):147, 2012.

[44] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[45] Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N16-1012`.

[46] Sumit Chopra, Michael Auli, Alexander M Rush, and SEAS Harvard. Abstractive sentence summarization with attentive recurrent neural networks. In *HLT-NAACL*, pages 93–98, 2016.

[47] James Clarke and Mirella Lapata. Global inference for sentence compression an integer linear programming approach. *J. Artif. Int. Res.*, 31(1):399–429, March 2008. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=1622655.1622667`.

[48] Arman Cohan and Nazli Goharian. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1045`.

[49] Arman Cohan and Nazli Goharian. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1.

[50] Arman Cohan and Nazli Goharian. Scientific document summarization via citation contextualization and scientific discourse. *International Journal on Digital Libraries*, pages 1–17, 2017. ISSN 1432-1300. doi: 10.1007/s00799-017-0216-8. URL http://dx.doi.org/10.1007/s00799-017-0216-8.

[51] Arman Cohan and Nazli Goharian. Contextualizing citations for scientific summarization using word embeddings and domain knowledge. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, 2017. ISBN 978-1-4503-5022-8/17/08. doi: 10.1145/3077136.3080740. URL http://doi.acm.org/10.1145/3077136.3080740.

[52] Arman Cohan, Luca Soldaini, and Nazli Goharian. Matching citation text and cited spans in biomedical literature: a search-oriented approach. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1042–1048, Denver, Colorado, May–June 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N15-1110.

[53] Arman Cohan, Luca Soldaini, Nazli Goharian, Allan Fong, Filice Ross, and Ratwani Raj. Identifying significance of discrepancies in radiology reports. In *SIAM International Conference on Data Mining (SDM) - Workshop on Data*

Mining for Medicine and Healthcare (DMMH), volume 5, pages 41–48, may 2016.

[54] Arman Cohan, Sydney Young, and Nazli Goharian. Triaging mental health forum posts. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 143–147, jun 2016.

[55] Arman Cohan, Allan Fong, Nazli Goharian, and Raj Ratwani. A neural attention model for categorizing patient safety events. In Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait, editors, *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*, pages 720–726, 2017. ISBN 978-3-319-56608-5. doi: 10.1007/978-3-319-56608-5_71. URL http://dx.doi.org/10.1007/978-3-319-56608-5_71.

[56] Arman Cohan, Allan Fong, Raj M. Ratwani, and Nazli Goharian. Identifying harm events in clinical care through medical narratives. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, pages 52–59, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4722-8. doi: 10.1145/3107411.3107485. URL http://doi.acm.org/10.1145/3107411.3107485.

[57] Arman Cohan, Sydney Young, Andrew Yates, and Nazli Goharian. Triaging content severity in online mental health forums. *Journal of the Association for Information Science and Technology (JASIST)*, 2017. doi: 10.1002/asi.23865. URL http://dx.doi.org/10.1002/asi.23865.

[58] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Change, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, 2018.

[59] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 2011. ISSN 1532-4435.

[60] John M Conroy and Hoa Trang Dang. Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 145–152. Association for Computational Linguistics, 2008.

[61] John M Conroy and Sashka T Davis. Vector space and language models for scientific document summarization. In *Proceedings of NAACL-HLT*, pages 186–191, 2015.

[62] John M Conroy, Judith D Schlesinger, Jeff Kubina, Peter A Rankel, and Dianne P OâĂŹLeary. Classy 2011 at tac: Guided and multi-lingual summaries and evaluation metrics. In *Proceedings of the Text Analysis Conference*, 2011.

[63] John M Conroy, Judith D Schlesinger, and Dianne P O'Leary. Nouveau-rouge: A novelty metric for update summarization. *Computational Linguistics*, 37(1): 1–8, 2011.

[64] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pages 51–60, 2014.

[65] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in twitter. In *ICWSM*, 2014.

[66] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *CLPysch*, pages 1–10, 2015.

[67] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado, June 5 2015. Association for Computational Linguistics.

[68] Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. *NAACL HLT 2015*, page 31, 2015.

[69] Glen Coppersmith, Kim Ngo, Ryan Leary, and Anthony Wood. Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 106–117, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[70] Glen Coppersmith, Casey Hilland, Ophir Frieder, and Ryan Leary. Scalable mental health analysis in the clinical whitespace via natural language processing. In *Biomedical & Health Informatics (BHI), 2017 IEEE EMBS International Conference on*, pages 393–396. IEEE, 2017.

[71] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.

[72] George E Dahl, Tara N Sainath, and Geoffrey E Hinton. Improving deep neural networks for lvcsr using rectified linear units and dropout. In *ICASSP*, pages 8609–8613. IEEE, 2013.

[73] Bo Dao, Thin Nguyen, Svetha Venkatesh, and Dinh Phung. Nonparametric discovery of online mental health-related communities. In *International Conference on Data Science and Advanced Analytics*, DSAA '15, pages 1–10. IEEE, 2015.

[74] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *International AAAI Conference on Web and Social Media*, ICWSM '14. AAAI, 2014.

[75] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 47–56, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1889-1. doi: 10.1145/2464464.2464480.

[76] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *ICWSM*, page 2, 2013.

[77] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI, 2013.

[78] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting Depression via Social Media. AAAI, jul 2013. URL `https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/`.

[79] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858207.

[80] Anita de Waard and Henk Pander Maat. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55. ACL, 2012.

[81] Anita De Waard and Henk Pander Maat. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 47–55. Association for Computational Linguistics, 2012.

[82] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[83] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91. Association for Computational Linguistics, 2011.

[84] Bart Desmet and Véronique Hoste. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

[85] Molla S Donaldson, Janet M Corrigan, Linda T Kohn, et al. *To err is human: building a safer health system*, volume 6. National Academies Press, 2000.

[86] Cícero Nogueira dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78, 2014.

[87] Mark Dredze. How social media will change public health. *IEEE Intelligent Systems*, 27(4):81–84, 2012.

[88] Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, August 2016. Association for Computational Linguistics.

[89] Johannes C Eichstaedt, Hansen Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, Sneha Jha, Megha Agrawal, Lukasz A Dziurzynski, Maarten Sap, et al. Psychological language on twitter predicts county-level heart disease mortality. *Psychological science*, 26(2):159–169, 2015.

[90] Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir Radev. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62, 2008.

[91] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2), 1990.

[92] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.(JAIR)*, 22(1):457–479, 2004.

[93] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[94] Manaal Faruqui, Jesse Dodge, Kumar Sujay Jauhar, Chris Dyer, Eduard Hovy, and A. Noah Smith. Retrofitting word vectors to semantic lexicons. In *NAACL-HLT*, pages 1606–1615. Association for Computational Linguistics, 2015. URL `http://aclweb.org/anthology/N15-1184`.

[95] Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics, 2010.

[96] Allan Fong, A Zachary Hettinger, and Raj M Ratwani. Exploring methods for identifying related patient safety events using structured and unstructured data. *Journal of biomedical informatics*, 58:89–95, 2015.

[97] Mark Garzone and Robert E Mercer. Towards an automated citation classifier. In *Conference of the Canadian Society for Computational Studies of Intelligence*, pages 337–346. Springer, 2000.

[98] George Giannakopoulos and Vangelis Karkaletsis. Summary evaluation: Together we stand npower-ed. In *Computational Linguistics and Intelligent Text Processing*, pages 436–450. Springer, 2013.

[99] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research (JAIR)*, 57:345–420, 2016.

[100] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM, 2001.

[101] Yvette Graham. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *EMNLP '15'*, pages 128–137, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1013`.

[102] Alex Graves. Supervised sequence labelling with recurrent neural networks. *Ph.D. thesis, Technische UniversitÂĺat MÂĺunchen*, 2008.

[103] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[104] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013.

[105] Annette M Green. Kappa statistics for multiple raters using categorical classifications. In *Proceedings of the 22nd annual SAS User Group International conference*, volume 2, page 4, 1997.

[106] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P16-1154`.

[107] Shengbo Guo and Scott Sanner. Probabilistic latent maximal marginal relevance. In *SIGIR*, pages 833–834. ACM, 2010.

[108] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *NAACL-HLT '09*, pages 362–370. Association for Computational Linguistics, 2009.

[109] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[110] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[111] Myriam Hernández-alvarez and José M Gomez. Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, 22 (03):327–349, 2016.

[112] William Hersh and Ellen Voorhees. Trec genomics special issue overview. *Information Retrieval*, 12(1):1–15, 2009. ISSN 1386-4564. doi: 10.1007/s10791-008-9076-6.

[113] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with genuine similarity estimation. *Comput. Linguist.*, 41(4):665–695, December 2015. ISSN 0891-2017. doi: doi:10.1162/COLI_a_00237. URL http://dx.doi.org/10.1162/COLI_a_00237.

[114] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[115] Kristy Hollingshead and Lyle Ungar, editors. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Association for Computational Linguistics, San Diego, California, USA, June 2016.

[116] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL https://aclweb.org/anthology/D/D15/D15-1162.

[117] Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 591–598, 2016.

[118] Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. Automated summarization evaluation with basic elements. In *LREC '06*, pages 604–611. Citeseer, 2006.

[119] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223. Association for Computational Linguistics, 2003.

[120] Samuel Huston and W Bruce Croft. Evaluating verbose query processing techniques. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 291–298. ACM, 2010.

[121] Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. Overview of the 2nd computational linguistics scientific document summarization shared task (cl-scisumm 2016). In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2016)*, 2016.

[122] John T. James. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care. *Journal of Patient Safety*, 9:122–128, 2013. ISSN 1549-8417. doi: 10.1097/PTS.0b013e3182948a69.

[123] Sue Jamison-Powell, Conor Linehan, Laura Daley, Andrew Garbett, and Shaun Lawson. I can't get no sleep: discussing# insomnia on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1501–1510. ACM, 2012.

[124] Rahul Jha, Reed Coke, and Dragomir Radev. Surveyor: A system for generating coherent survey articles for scientific topics. *Ann Arbor*, 1001:48109, 2015.

[125] Fanghong Jian, Jimmy Xiangji Huang, Jiashu Zhao, Tingting He, and Po Hu. A simple enhancement for ad-hoc information retrieval via topic modelling. In *SIGIR*, pages 733–736. ACM, 2016.

[126] Hongyan Jing. Using hidden markov modeling to decompose human-written summaries. *Computational linguistics*, 28(4):527–543, 2002.

[127] Eamon Johnson, W Christopher Baughman, and Gultekin Ozsoyoglu. Mixing domain rules with machine learning for radiology text classification. 2014.

[128] Adam N Joinson and Carina B Paine. Self-disclosure, privacy and the internet. *Oxford handbook of Internet psychology*, pages 237–252, 2007.

[129] K Sparck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information Processing & Management*, 36(6):809–840, 2000.

[130] Natalie J Jones and Craig Bennell. The development and validation of statistical prediction rules for discriminating between genuine and simulated suicide notes. *Archives of Suicide Research*, 11(2):219–233, 2007.

[131] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[132] David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. Citation classification for behavioral analysis of a scientific field. *CoRR*, 2016.

[133] Amit D Kalaria and Ross W Filice. Comparison-bot: an automated preliminary-final report comparison system. *Journal of digital imaging*, pages 1–6, 2015.

[134] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 655–665. Association for Computational Linguistics, 2014. URL http://aclweb.org/anthology/P14-1062.

[135] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. In *ACL*, pages 655–665, 2014.

[136] Debra L Karch, Linda L Dahlberg, Nimesh Patel, Terry W Davis, Joseph E Logan, Holly A Hill, L Ortega, et al. Surveillance for violent deathsâĂŤnational violent death reporting system, 16 states, 2006. *MMWR Surveill Summ*, 58(1): 1–44, 2009.

[137] Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. Utilizing context in generative bayesian models for linked corpus. In *AAAI*, volume 10, page 1, 2010.

[138] Sunghwan Mac Kim, Yufei Wang, Stephen Wan, and Cecile Paris. Data61-csiro systems at the clpsych 2016 shared task. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 128–132, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[139] Yoon Kim. Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751, 2014.

[140] JP Kincaid, RP Fishburne, RL Rogers, and BS Chissom. Derivation of new readability formulas. Technical report, Technical report, TN: Naval Technical Training, US Naval Air Station, Memphis, TN, 1975.

[141] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[142] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprints*, arXiv:1412.6980, 2014.

[143] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. In *NIPS*, Cambridge, MA, USA, 2015. MIT Press.

[144] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.

[145] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. *Comput. Linguist.*, 32(4), December 2006. ISSN 0891-2017.

[146] Stefan Klampfl, Andi Rexha, and Roman Kern. Identifying referenced text in scientific publications by summarisation and classification techniques. In *BIRNDL at JCDL*, pages 122–131, 2016.

[147] Louisa Lam and SY Suen. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(5):553–568, 1997.

[148] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[149] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[150] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.

[151] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86 (11), 1998.

[152] David Lester. The final hours: A linguistic analysis of the final words of a suicide. *Psychological reports*, 106(3):791–797, 2010.

[153] Lei Li, Liyuan Mao, Yazhao Zhang, Junqi Chi, Taiwen Huang, Xiaoyue Cong, and Heng Peng. Cist system for cl-scisumm 2016 shared task. In *BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries*, 2016.

[154] Piji Li, Wai Lam, Lidong Bing, Weiwei Guo, and Hang Li. Cascaded attention based unsupervised information distillation for compressive summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2071–2080, 2017.

[155] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, 2004.

[156] Jimmy Lin, Nitin Madnani, and Bonnie J Dorr. Putting the user in the loop: interactive maximal marginal relevance for query-focused summarization. In *NAACL-HLT*, pages 305–308. Association for Computational Linguistics, 2010.

[157] Jeffrey Ling and Alexander Rush. Coarse-to-fine attention models for document summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 33–42, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W17-4505`.

[158] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[159] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[160] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39, 2016.

[161] Annie Louis and Ani Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *IJCNLP*, pages 605–613, 2011.

[162] Annie Louis and Ani Nenkova. Verbose, laconic or just right: A simple computational model of content appropriateness under length constraints. *EACL 2014*, page 636, 2014.

[163] Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[164] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[165] Carl Macrae. The problem with incident reporting. *BMJ Quality & Safety*, 25: 71–75, 2016. ISSN 2044-5415. doi: 10.1136/bmjqs-2015-004732.

[166] Martin A Makary and Michael Daniel. Medical errorâĂŤthe third leading cause of death in the us. *Bmj*, 353:i2139, 2016.

[167] Shervin Malmasi, Marcos Zampieri, and Mark Dras. Predicting post severity in mental health forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 133–137, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[168] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL `http://www.aclweb.org/anthology/P/P14/P14-5010`.

[169] G Harry McLaughlin. Smog grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[170] Yishu Miao and Phil Blunsom. Language as a latent variable: Discrete generative models for sentence compression. *arXiv preprint arXiv:1609.07317*, 2016.

[171] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[172] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[173] Doug Millen. Reachout annual report 2013/2014. 2015. URL `http://about.au.reachout.com/us/annual-reports-financials`.

[174] George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[175] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. Clpsych 2016 shared task: Triaging content in online peer-support forums. *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–127, jun 2016.

[176] David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 106–117, San Diego, CA, USA, June 2016. Association for Computational Linguistics.

[177] Imogen Mitchell, Anne Schuster, Katherine Smith, Peter Pronovost, and Albert Wu. Patient safety incident reporting: a qualitative study of thoughts and perceptions of experts 15 years after âĂŸ To Err is Human '. *BMJ Quality & Safety*, 25:92–99, 2016. ISSN 2044-5415. doi: 10.1136/bmjqs-2015-004405.

[178] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, 2015.

[179] Luis Moraes, Shahryar Baki, Rakesh Verma, and Daniel Lee. University of houston at cl-scisumm 2016: Svms with tree kernels and sentence similarity. In *BIRNDL@ JCDL*, pages 113–121, 2016.

[180] Danielle Mowery, Albert Park, Mike Conway, and Craig Bryan. Towards automatically classifying depressive symptoms from twitter data for population

health. In *Proceedings of the Workshop on Computational Modeling of PeopleâĂŹs Opinions, Personality, and Emotions in Social Media*, pages 182–191, 2016.

[181] Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *NAACL-HLT*, 2016.

[182] Elizabeth L Murnane and Scott Counts. Unraveling abstinence and relapse: smoking cessation reflected in social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 1345–1354. ACM, 2014.

[183] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Johannes FÃijrnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010.

[184] Preslav I Nakov, Ariel S Schwartz, and Marti Hearst. Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR'04 workshop on Search and Discovery in Bioinformatics*, pages 81–88, 2004.

[185] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[186] National Center for Injury Prevention and Control. Suicide facts at a glance. In *Centers for Disease Control and Prevention*, 2015. URL `https://www.cdc.gov/violenceprevention/pdf/suicide-datasheet-a.pdf`.

[187] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 2004. URL `http://aclweb.org/anthology/N04-1019`.

[188] Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4, 2007.

[189] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[190] MEJ Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1):25–31, 2012.

[191] Dung HM Nguyen and Jon D Patrick. Supervised machine learning and active learning in classification of radiology reports. *Journal of the American Medical Informatics Association*, 21(5):893–901, 2014.

[192] Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Jordi Alonso, Matthias Angermeyer, Annette Beautrais, Ronny Bruffaerts, Wai Tat Chiu, Giovanni De Girolamo, Semyon Gluzman, et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. *The British Journal of Psychiatry*, 192(2):98–105, 2008.

[193] Tadashi Nomoto. Neal: A neurally enhanced approach to linking citation and reference. In *BIRNDL 2016 Joint Workshop on Bibliometric-enhanced Information Retrieval and NLP for Digital Libraries*, 2016.

[194] Alexandra Olteanu, Onur Varol, and Emre Kıcıman. Towards an open-domain framework for distilling the outcomes of personal experiences from social media timelines. In *Proceedings of International AAAI Conference on Web and Social Media*, ICWSM '16'. AAAI, May 2016.

[195] Mei-Sing Ong, Farah Magrabi, and Enrico Coiera. Automated identification of extreme-risk events in clinical incident reports. *Journal of the American Medical Informatics Association*, 19(e1):e110–e118, 2012.

[196] David Opitz and Richard Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[197] Miles Osborne. Using maximum entropy for sentence extraction. In *Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4*, pages 1–8. Association for Computational Linguistics, 2002.

[198] Karolina Owczarzak and Hoa Trang Dang. Overview of the tac 2011 summarization track: Guided task and aesop task. In *TAC 2011*, 2011.

[199] Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics, 2012.

[200] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL '02*, pages 311–318. Association for Computational Linguistics, 2002.

[201] Minsu Park, David W McDonald, and Meeyoung Cha. Perception differences between the depressed and non-depressed users in twitter. In *Proceedings of International AAAI Conference on Web and Social Media (ICWSM)*, 2013.

[202] Jon Parker, Andrew Yates, Nazli Goharian, and Ophir Frieder. Health-related hypothesis generation using social media data. *Social Network Analysis and Mining*, 5(1):1–15, 2015. ISSN 1869-5469. doi: 10.1007/s13278-014-0239-8. URL `http://dx.doi.org/10.1007/s13278-014-0239-8`.

[203] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *ICML*, 28:1310–1318, 2013.

[204] Michael Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *EMNLP*, pages 66–76. Association for Computational Linguistics, 2010. URL `http://aclweb.org/anthology/D10-1007`.

[205] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *International Conference On Web And Social Media (ICWSM)*, 20:265–272, 2011.

[206] Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. Social media mining for public health monitoring and surveillance. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 21, page 468, 2016.

[207] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.

[208] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*, 2015.

[209] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *EMNLP*, 12:1532–1543, 2014.

[210] Bethany Percha, Houssam Nassif, Jafi Lipson, Elizabeth Burnside, and Daniel Rubin. Automatic classification of mammography reports by bi-rads breast tissue composition class. *Journal of the American Medical Informatics Association*, 19(5):913–916, 2012.

[211] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, Kevin B Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(Suppl. 1):3, 2012.

[212] Emily Pitler, Annie Louis, and Ani Nenkova. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the ACL 2010*, pages 544–554. Association for Computational Linguistics, 2010. URL `http://aclweb.org/anthology/P10-1056`.

[213] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[214] Kenneth Portier, Greta E Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, et al. Understanding topics and sentiment in an online cancer survivor community. *JNCI Monographs*, 47:195–198, 2013.

[215] Martin Potthast, Matthias Hagen, Michael Völske, and Benno Stein. Crowdsourcing interaction logs to understand text reuse from the web. In *ACL (1)*, pages 1212–1221, 2013.

[216] Vahed Qazvinian and Dragomir R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 689–696, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6.

[217] Vahed Qazvinian and Dragomir R Radev. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 555–564. Association for Computational Linguistics, 2010.

[218] Vahed Qazvinian, DR Radev, and SM Mohammad. Generating Extractive Summaries of Scientific Paradigms. *J. Artif. Intell. Res.(...*, 46:165–201, 2013.

[219] Vahed Qazvinian, Dragomir R Radev, Saif Mohammad, Bonnie J Dorr, David M Zajic, Michael Whidby, and Taesun Moon. Generating extractive summaries of scientific paradigms. *J. Artif. Intell. Res.(JAIR)*, 46:165–201, 2013.

[220] Peter Rankel, John M. Conroy, Eric V. Slud, and Dianne P. O'Leary. Ranking human and machine summarization systems. EMNLP '11, pages 467–473,

Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4. URL http://dl.acm.org/citation.cfm?id=2145432.2145486.

[221] Raj Ratwani, Allan Fong, Ross Filice, Arman Cohan, Luca Soldaini, Nazli Goharian, and Ophir Frieder. Systems and methods for targeted radiology resident training, July 2017. US Patent App. 15/410,850.

[222] Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression. In *EMNLP*, pages 1348–1353. Association for Computational Linguistics, 2013.

[223] Anthony Rios and Ramakanth Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *ACM-BCB*, pages 258–267. ACM, 2015.

[224] Anna Ritchie, Stephen Robertson, and Simone Teufel. Comparing citation contexts for information retrieval. In *CIKM*, pages 213–222. ACM, 2008. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458113. URL http://doi.acm.org/10.1145/1458082.1458113.

[225] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[226] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

[227] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL `http://aclweb.org/anthology/D15-1044`.

[228] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

[229] Alexander T Ruutiainen, Mary H Scanlon, and Jason N Itri. Identifying benchmarks for discrepancy rates in preliminary interpretations provided by radiology trainees at an academic institution. *Journal of the American College of Radiology*, 8(9):644–648, 2011.

[230] Alexander T Ruutiainen, Daniel J Durand, Mary H Scanlon, and Jason N Itri. Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight. *Academic radiology*, 20(3): 305–311, 2013.

[231] Samuli I Saarni, Jaana Suvisaari, Harri Sintonen, Sami Pirkola, Seppo Koskinen, Arpo Aromaa, and JOUKO LÖNNQVIST. Impact of psychiatric disorders on health-related quality of life: general population survey. *The British journal of psychiatry*, 190(4):326–332, 2007.

[232] Horacio Saggion, Ahmed AbuRaâĂŹed, and Francesco Ronzano. Trainable citation-enhanced summarization of scientific articles. In *Cabanac G, Chandrasekaran MK, Frommholz I, Jaidka K, Kan M, Mayr P, Wolfram D, editors. Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL);*

2016 June 23; Newark, United States. CEUR Workshop Proceedings:[Sl]; 2016. p. 175-86. CEUR Workshop Proceedings, 2016.

[233] Ágnes Sándor and Anita De Waard. Identifying claimed knowledge updates in biomedical research articles. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 10–17. ACL, 2012.

[234] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[235] H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W14-3214`.

[236] Abigail See, Christopher Manning, and Peter Liu. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*, 2017. URL `https://arxiv.org/abs/1704.04368`.

[237] Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219, 2017.

[238] Richard E Sharpe Jr, David Surrey, Richard JT Gorniak, Levon Nazarian, Vijay M Rao, and Adam E Flanders. Radiology report comparator: a novel

method to augment resident education. *Journal of digital imaging*, 25(3):330–336, 2012.

[239] Morton M Silverman and Ronald W Maris. The prevention of suicidal behaviors: An overview. *Suicide and Life-Threatening Behavior*, 25(1):10–21, 1995.

[240] CT Snomed. Systematized nomenclature of medicine-clinical terms. *International Health Terminology Standards Development Organisation*, 2011.

[241] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231, 2006.

[242] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[243] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[244] Jacopo Staiano and Marco Guerini. Depeche mood: a lexicon for emotion analysis from crowd annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 427–433, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-2070`.

[245] Jacopo Staiano and Marco Guerini. Depechemood: a lexicon for emotion analysis from crowd-annotated news. *arXiv preprints*, arXiv:1405.1605, 2014.

[246] Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. Overview of the pan/clef 2015 evaluation lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 518–538. Springer, 2015.

[247] Barbara Starfield. Is us health really the best in the world? *Jama*, 284(4): 483–485, 2000.

[248] Josef Steinberger and Karel Jezek. Using latent semantic analysis in text summarization and summary evaluation. In *Proc. ISIM'04*, pages 93–100, 2004.

[249] Josef Steinberger and Karel Jezek. Using latent semantic analysis in text summarization and summary evaluation. In *ISIM*, pages 93–100, 2004.

[250] Tara W Strine, Ali H Mokdad, Lina S Balluz, Olinda Gonzalez, Raquel Crider, Joyce T Berry, and Kurt Kroenke. Depression and anxiety in the united states: findings from the 2006 behavioral risk factor surveillance system. *Psychiatric Services*, 2015.

[251] Frederick Suppe. The structure of a scientific paper. *Philosophy of Science*, 65 (3):381–405, 1998.

[252] Simon Šuster, Stéphan Tulkens, and Walter Daelemans. A short review of ethical challenges in clinical natural language processing. *arXiv preprint arXiv:1703.10090*, 2017.

[253] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[254] Barbara G Tabachnick, Linda S Fidell, and Steven J Osterlind. Using multivariate statistics. 2001.

[255] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *EMNLP*, pages 1422–1432, 2015.

[256] Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Comput. Linguist.*, 28(4):409–445, December 2002. ISSN 0891-2017. doi: 10.1162/089120102762671936. URL `http://dx.doi.org/10.1162/089120102762671936`.

[257] Simone Teufel and Marc Moens. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445, 2002.

[258] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. Automatic classification of citation function. *EMNLP '06*, page 103, 2006.

[259] Paul Thompson, Craig Bryan, and Chris Poulin. Predicting military and veteran suicide risk: Cultural aspects. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–6, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

[260] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3187–3196. ACM, 2015.

[261] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[262] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

[263] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

[264] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam Med*, 37(5):360–363, 2005.

[265] Jessica Walls, Natalie Hunter, Penelope MA Brasher, and Stephen GF Ho. The depictors study: discrepancies in preliminary interpretation of ct scans between on-call residents and staff. *Emergency radiology*, 16(4):303–308, 2009.

[266] Sida Wang and Christopher D Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 90–94. Association for Computational Linguistics, 2012.

[267] Huw Williams, Alison Cooper, and Andrew Carson-Stevens. Opportunities for incident reporting. *BMJ Quality & Safety*, 25:133–134, 2016. ISSN 2044-5415.

[268] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[269] Sam Wiseman, Stuart M Shieber, and Alexander M Rush. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*, 2017.

[270] Yijun Xiao and Kyunghyun Cho. Efficient character-level document classification by combining convolution and recurrent layers. *arXiv preprints*, arXiv:1602.00367, 2016.

[271] Haotian Xu, Ming Dong, Dongxiao Zhu, Alexander Kotov, April Idalski Carcone, and Sylvie Naar-King. Text classification with topic-based word embedding and convolutional neural networks. In *ACM-BCB*, pages 88–97, 2016.

[272] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL-HLT*, pages 1480–1489, 2016.

[273] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D17-1322`.

[274] Antonio Jose Jimeno Yepes, James G Mork, Dina Demner-Fushman, and Alan R Aronson. Comparison and combination of several mesh indexing approaches. In *AMIA annual symposium proceedings*, volume 2013, page 709, 2013.

[275] Qing Zeng-Treitler, Long Ngo, Sasikiran Kandula, Graciela Rosemblat, Hyeon-Eui Kim, and Brent Hill. A method to estimate readability of health content. *Association for Computing Machinery*, 2012.

[276] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342. ACM, 2001. ISBN 1-58113-331-6. doi: 10.1145/383952.384019. URL `http://doi.acm.org/10.1145/383952.384019`.

[277] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.