Paper 1264-2017

Mind-map the Gap – Sentiment Analysis of Public Transport

Tamás Bosznay, Amadeus Software Ltd.

ABSTRACT

This paper presents a case study where social media posts by individuals related to public transport companies in the United Kingdom are collected from social media sites such as Twitter using SAS®. The posts are then further processed in SAS® Text Miner to retrieve brand names, means of public transport (underground, trains, buses) and their attributes mentioned.

Relevant concepts and topics are identified using text mining techniques and visualized using concept links and word clouds.

Later we aim to identify and categorize sentiments against public transport in the corpus of the posts.

Finally, we approach to create an association map/mind-map of the different service dimensions/topics and the brands of public transport using correspondence analysis.

INTRODUCTION

Daily commuting in cities can be a tough job. People usually form an opinion about the transport services they use - these sentiments may be negative or show satisfaction with the services. People not only form their opinion – but in the age of social media and smartphones they frequently share it with the public – via Twitter, Facebook or even forum posts. Allowing this way of sharing opinions and sentiments are beneficial for the public transport authorities and companies via giving the ability for their customers to release some steam. Also, this way the patterns customers reveal through their opinions might be used for planning e.g. altering services or introducing extra customer communication to increase customer satisfaction.

“Mind the gap” – the iconic station safety announcement originated from London Underground became a popular slogan globally since it has been introduced in 1969. Mind maps are commonly used both professionally and privately to visually organize information. Trying to merge these two phrases does not seem to make much sense – however, we aim to achieve this in our paper. We will create mind maps for the gap in public transport – in this sense the gap could be interpreted as a main driver for a customer of public transport to share their thoughts and feelings about it (caused by the gap between their expectations and their experience).

In our paper, we aim to show techniques using SAS® to collect and mine these opinions and sentiments from social media. We extensively use the capabilities of SAS® Text Miner and Base SAS® (via Enterprise Guide), both in collecting and accessing the raw data and analyzing/visualizing it. We also use SAS® Visual Analytics to create some specific visualizations.

METHODOLOGY

The steps of the analysis process followed are shown below:

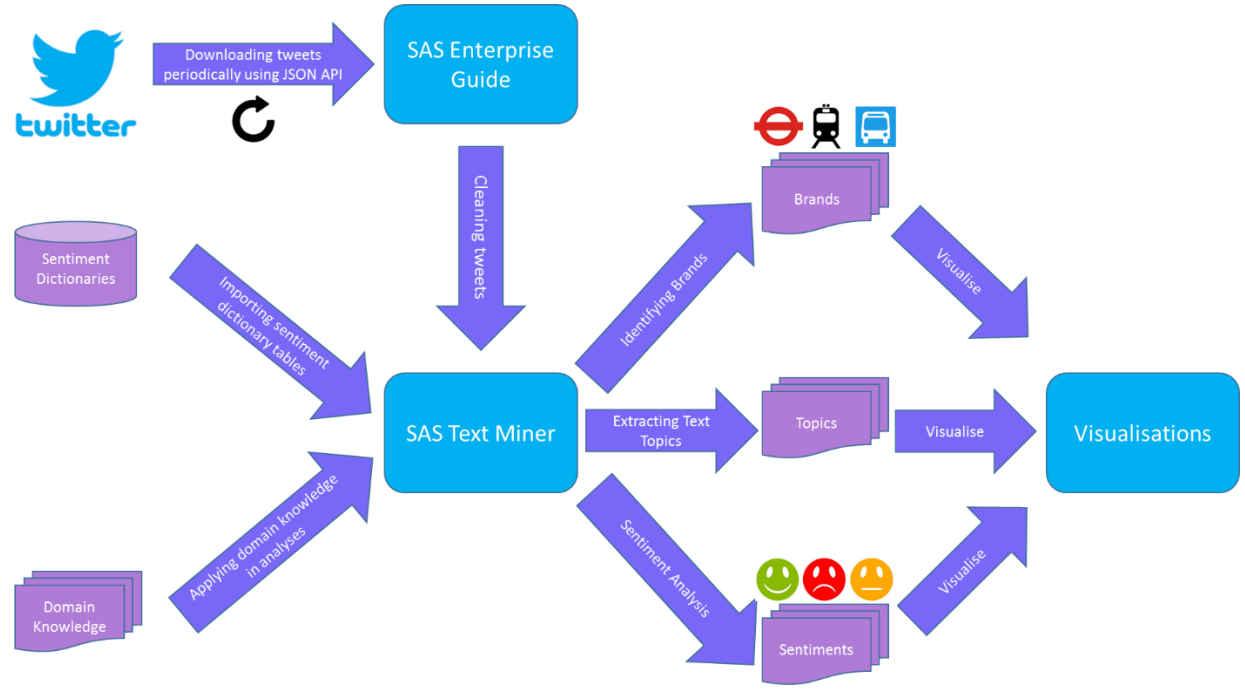


Figure 1: Analysis process followed

As the diagram shows, first the options to access social media data are identified and discussed. After downloading data using the preferred method, some preparation and cleaning steps are done to be able to use the data in SAS® Text Miner (Data Collection Phase). External information such as sentiment dictionaries and domain knowledge is used for augmenting the language pack included in SAS® Text Miner. The actual analytics steps are aiming to identify brands, extract topics and perform sentiment analysis on the data (Text Analytics Phase). Finally, results are visualised and conclusions are drawn.

DATA COLLECTION PHASE

EXPLORING RELEVANT CONTENT IN SOCIAL MEDIA

As we would like to collect and analyse social media data related to public transport, we first need to look after sources (Facebook fan pages, tweets from Twitter) where the volume of comments/posts are large enough. As we are focusing on the United Kingdom, the obvious choice in terms of initial discussion area for large volumes is either anything about public transport in London, or about the whole of the country (rail looks like a good candidate for this). We have chosen to go for the city public transport as we believe it is generating much more discussion than rural or long-distance travel, also we can assume more variety of things influencing social media posts about city public transport and therefore more variety in the actual comments/tweets.

Looking at Facebook fan pages, it is relatively hard to find a considerably large corpus of comments on public transport in the United Kingdom. Twitter is much more actively used and therefore will be the source of data for our analyses. In other countries/markets, for example in Eastern Europe, Facebook is also actively used for posting short opinions and impressions about a wide range of discussion topics, including public transport.

Focusing on Twitter, after some initial exploration of the tweets available, we decided to take the approach to use Twitter's Search API to collect tweets that were posted with certain hashtags. The mostly used hashtag about public transport in London is the hashtag #tfl, which stands for Transport for London, the city-owned organisation for coordinating public transport. People also tend to use other hashtags when tweeting about their impressions related to public transport and commuting in London, like #londonunderground. The hashtag #tfl is far more the most actively used, and therefore we basically focused on collecting tweets using it.

ACCESSING DATA FROM SOCIAL MEDIA

SAS® Visual Analytics gives an easy-to-use facility to import data from Twitter in the Data Explorer, which uses Twitter's Search API. The search term could be easily provided and after authenticating against Twitter with a normal user ID, the downloading of the tweets starts. Twitter APIs, however, does not allow to search retrospectively for a long period of time. The Twitter results retrieved in SAS® Visual Analytics tend to return tweets related to the given search term only a couple of days back in time. For our analysis, we need a longer track record than that. This means we have to run the Twitter data collection periodically. In SAS® Visual Analytics, automating this data collection process is not a very straightforward task. Therefore, we decided to use Base SAS® to connect to the Twitter Search API and gather tweets.

This approach uses Twitter APIs not as SAS® Visual Analytics, but as a new, different application. Therefore, a different type of access is required. We must register our own Twitter App to be able to use the authentication method used with twitter APIs. This authentication method is known as OAuth 2. The method basically requires two types of credentials (a Consumer Key and another key, called Consumer Secure) to be provided to Twitter. As a response to this request, a Bearer Token is passed back to the requestor. The Bearer Token must be provided in the header of the actual search query request. The API is basically a RESTful API using JSON format. This means the response for the search query request will be in JSON format, it is therefore needed to have a method to parse the JSON results to be able to use them in SAS®.

This could be achieved either using data step or PROC DS2 in various SAS® versions. Alternatively, PROC GROOVY could be of good use. However, as SAS® 9.4 (TS1M4) has native JSON engine support, this approach has been used to parse the results. The engine works in a similar way to the XML libname engine with the automap option. If no special processing is needed, the basic JSON structure is parsed to have all the different entities in different tables.

To authenticate against Twitter, the method described in Reference [1] has been used. The following snippet shows PROC HTTP setup for getting and parsing the bearer token:

```
%let CONSUMER_KEY=<<your consumer key>>;
%let CONSUME_SECRET=<<your consumer secret>>;
%let BEARER_TOKENFILE=<<path where the bearer token file would be stored permanently >>;

/* create temp files for the content and header input streams */
filename intoken TEMP lrecl=2048;
filename hdrint TEMP lrecl=2048;

 /* keep the responses permanently */
 filename outtoken "&BEARER_TOKENFILE.";

 /* post request content is the grant_type */
 data _null_;
   file intoken;
   put "grant_type=client_credentials&";
 run;

 /* request the bearer token by providing consumer key and secret */
 data _null_;
   file hdrint;
   consumerKey = urlencode("&CONSUMER_KEY.");
   consumerSecret = urlencode("&CONSUME_SECRET.");
   encodedAccessToken  = put( compress(consumerKey || ":" ||
consumerSecret),$base64x32767.);
   put "Authorization: Basic " encodedAccessToken;
 run;
```

```
 /* sending the request */
 proc http method="post"
    in=intoken out=outtoken
    headerin=hdrint
    url="https://api.twitter.com/oauth2/token"
    ct="application/x-www-form-urlencoded;charset=UTF-8";
 run;

/* parsing the authentication request response file */
 libname libtoken json "&BEARER_TOKENFILE.";

/* getting the Bearer Token */
 data libtokenalldata;
 set libtoken.alldata;
 if p1='access_token' then
 call symputx('BEARER_TOKEN',value);
 run;
```

The next step is to put the actual search API query request together and send it to Twitter:

```
%let TWITTER_RESULTS=<<path where the query results file would be stored permanently >>;
%let TWITTER_QUERY=%23tfl;

 /* create temp file for the header input stream */
filename hdrin TEMP lrecl=2048;

 /* keep the responses permanently */
filename out "&TWITTER_RESULTS.";

/* write required syntax in the header*/
data _null_;
    file hdrin;
    put "Authorization: Bearer &BEARER_TOKEN.";
run;

 /* searching for tweets, limiting the number of results to 10 */
 proc http method="get"
    out=out headerin=hdrin
    url="https://api.twitter.com/1.1/search/tweets.json?q=&TWITTER_QUERY.&count=10"
    ct="application/x-www-form-urlencoded;charset=UTF-8";
 run;
```

The final step is to use the JSON libname engine to parse the results in SAS®:

```
libname outcome json "&TWITTER_RESULTS.";
```

PREPARING SOCIAL MEDIA DATA FOR ANALYTICS

The tables generated by parsing the Twitter Search API results is the following structure:

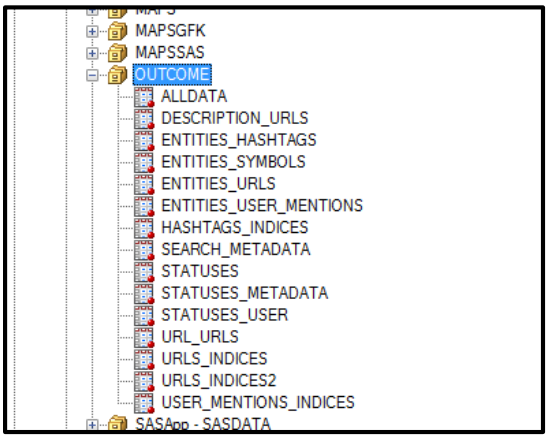


Figure 2: Result tables from Twitter Search JSON API

Having the content of the tables shows we need to use the STATUSES table for our analysis. To have user information (user names, number of followers etc.) we need to join it to the STATUSES_USER table. We will create a table for our analyses having the following column structure:

Name	Type	Description
created_at	Character	Date and Time of the tweet
id	Numeric	Tweet ID
text	Character	Tweet text
source	Character	Website or app the tweet was tweeted from
retweet_count	Numeric	Number of retweets on the tweet
retweeted	Numeric	Whether the tweet was retweeted
name	Character	Twitter User Name (Free format)
screen_name	Character	Twitter User Screen Name (the @name)
location	Character	Location of user
description	Character	Description of user
url	Character	Url in the tweet
followers_count	Numeric	Number of followers of the user
friends_count	Numeric	Number of friends of the user
statuses_count	Numeric	Number of tweets by the user

Table 1: Tweets base table structure

Now, as we need to build a track record of fair volume of tweets via running this collection periodically, we are appending the results to build our base table. As some tweets might be returned repeatedly in different runs, it is needed to deduplicate the data by Tweet ID.

Alternatively, we could use the Twitter Streaming API, especially if more real-time analysis would be in our scope. As we are doing retrospective analysis, we stick to the API usage described above.

Social Media APIs tend to change from time to time. The OAuth 2 authentication method will be used for a longer time than the API results structure remains fixed. Therefore, parsing Twitter query results needs a check step built in to ensure the API results structure is unchanged.

As a result of running these data collection steps, about 4,200 unique tweets have been stored for the period 8 November 2016 to 8 February 2016 by searching for the hashtag #tfl.

From this point onwards SAS® Enterprise Guide®, SAS® Enterprise Miner™ (Text Miner) and SAS® Visual Analytics will be used as clients to commence with the analyses.

TEXT ANALYTICS PHASE

INFORMATION RETRIEVAL TO IDENTIFY BRANDS

Now we have our base table of tweets. There are a couple of considerations specific to Twitter data and its usage in text analytics:

- Tweets can be retweets of other tweets. This can be identified by pattern searching for “Retweeted” in the text field. Removing these resulted in 4,084 tweets in our table.
- Tweets can tag twitter users using the @screen_name pattern. This might mean relevant data for text analytics as the tagging might happen to mention the name of another user, who is also a relevant party in the actual area being tweeted about. Consequently, these user names should not necessarily be get rid of from the text field. However, SAS® Text Miner will flag them as entities (mainly as “Person”) after automatically getting rid of the @ special character, and later they might be used in the different text analytics steps.
- Tweets usually contain one or multiple hashtags in their text – hashtags were also our main search criteria. Text Miner will get rid of the leading # character from them, and handle them usually as Proper Nouns as most of them have no meaning. We still would like to keep them in our analyses because they carry information about the actual tweets. Alternatively, they can be handled by replacing # with a different prefix (e.g. “hashtag_”) to flag that they are hashtags. The results table from parsing the Twitter Search API JSON response has an Entities_Hashtags table which also lists all the hashtags (see Figure 3).
- Tweets frequently contain links. These are not required, so we are using Text Miner to get rid of them from the analyses via filtering out specific entities (see Figure 4).

As per our process diagram on Figure 1, our first aim is to identify the brands and initiatives associated with London public transport. We need to explore the tweets generally to get an idea about what brands are getting mentioned or tweeted about and review those in line with what we expect.

The easiest way to approach this is to start by extracting terms and entities that can be related to public transport using our tweets table. To achieve this, we set up the following simple Enterprise Miner diagram:

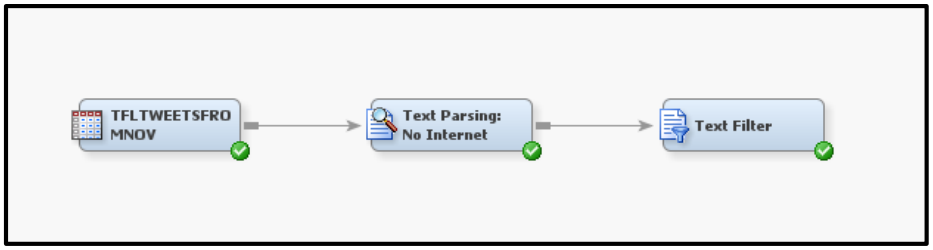


Figure 3: First Text Miner diagram to extract terms

We are changing the default settings in the Text Parsing nodes to ignore the links as discussed above. This can be one via setting “Find Entities” to “Standard” and selecting “Internet” in “Ignore Types of Entities” in the node properties. Also, depending our next steps, it might be beneficial to set the Text Parsing node not to detect different parts of speech – especially for creating visualisations.

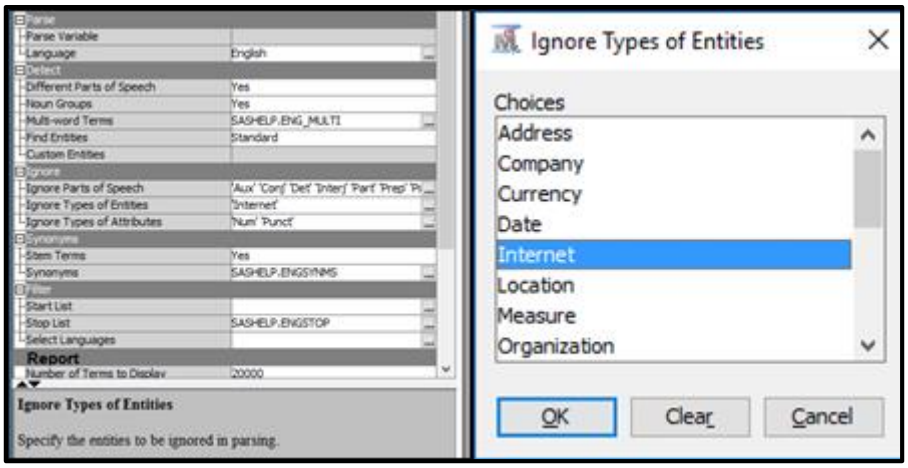
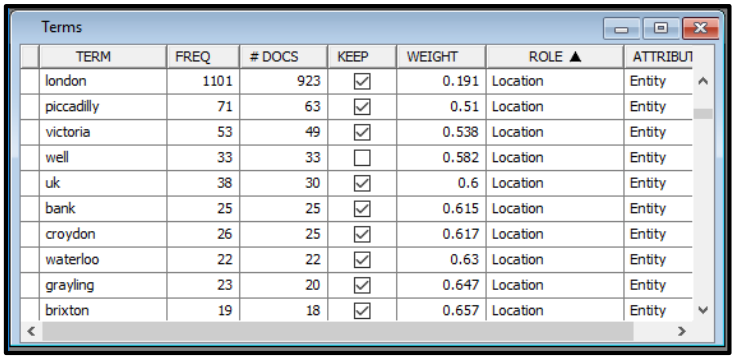


Figure 4: Ignoring Internet entities in the Text Parsing node

Using the settings on Figure 4 we run the diagram path and open the Interactive Filter Viewer in the Text Filter node properties. Now we can achieve the following:

- Examine the terms and entities presented
- Identify the terms and entities that might need to be handled as synonyms in our analyses
- Based on some expert judgement, keep (or drop) terms or entities that Text Miner suggested to drop (or keep) based on weights
- Identify brands related to public transport occurring in our tweets, based on the above

An example is shown below on Figure 5 to examine all the locations (Location type Entities) in our tweets.



TERM	FREQ	# DOCS	KEEP	WEIGHT	ROLE ▲	ATTRIBUT
london	1101	923	☑	0.191	Location	Entity
piccadilly	71	63	☑	0.51	Location	Entity
victoria	53	49	☑	0.538	Location	Entity
well	33	33	☐	0.582	Location	Entity
uk	38	30	☑	0.6	Location	Entity
bank	25	25	☑	0.615	Location	Entity
croydon	26	25	☑	0.617	Location	Entity
waterloo	22	22	☑	0.63	Location	Entity
grayling	23	20	☑	0.647	Location	Entity
brixton	19	18	☑	0.657	Location	Entity

Figure 5: Example set of Location entities

Another useful technique is to use the concept links in the Interactive Filter Viewer. The next example shows exploring the terms related to “tube” by gradually expanding the concept link diagram. This method helps us finding the relevant terms that can be used to determine the brand of a tweet.

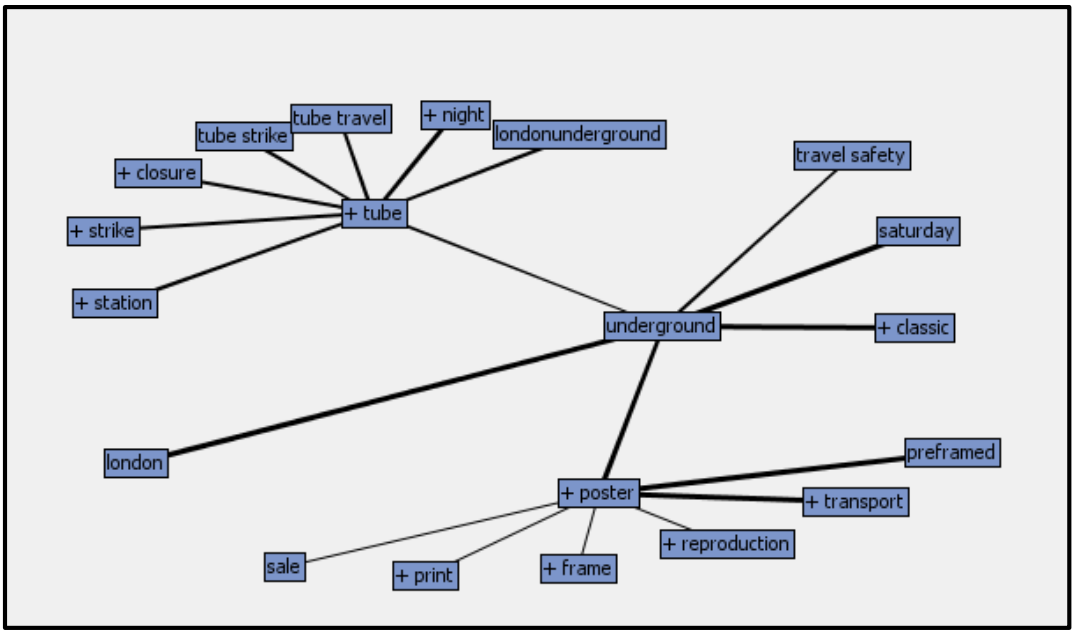


Figure 6: Gradually expanding the concept link diagram

As Figure 6 shows, in this initial exploration we did not consider “tube” and “underground” as synonyms, however, later it can be beneficial for our analyses (and similar logic to other cases of synonyms).

Using these exploration techniques and our domain knowledge, we came up with user created topics to identify the brands of the tweets. To achieve this, we add a Text Topic node and a Save Data node to further use the brands flags:

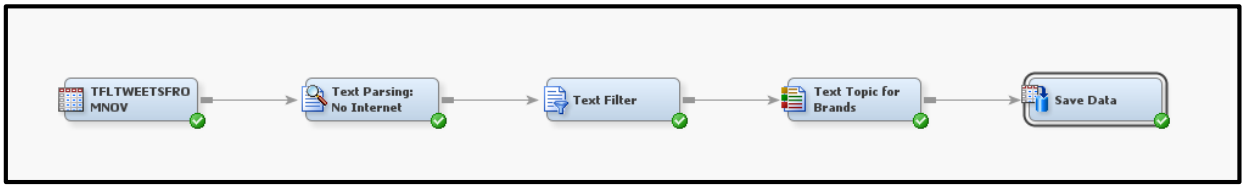


Figure 7: Diagram for flagging brands and saving brand flags data

In our Text Topic node, we would like to flag only the user topics on the document collection of tweets and not to have other topics (we will determine the conversations topics later, now we focus on brands). To achieve this, we set the Number of Multi-Term Topics from the default 25 to 0. Also, we upload our user topics table:

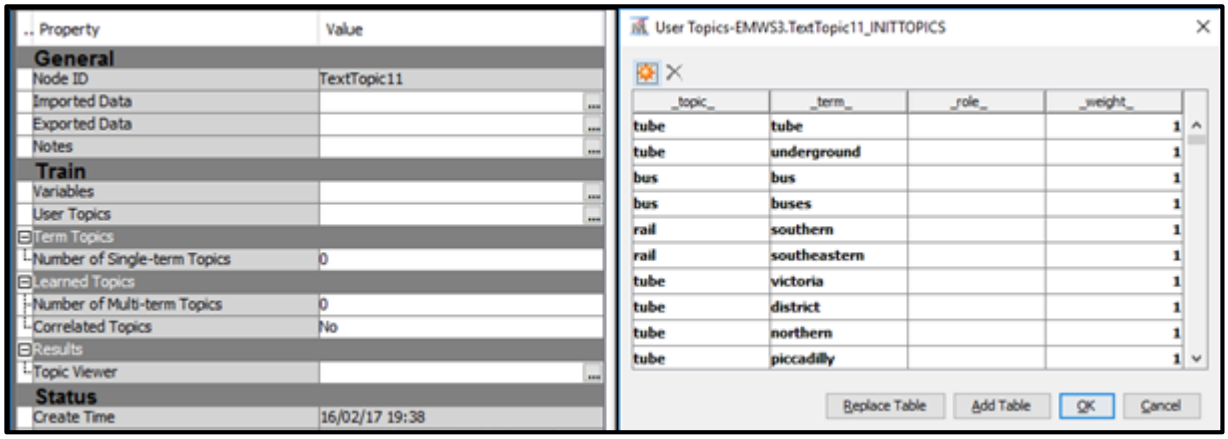


Figure 8: User topics data set to identify brands

This results in the Text Topic node generating flags for the user topics (the brands we specified).

As an alternative to this rule-based method of flagging the brands of the tweets, we could follow a supervised learning method. That includes flagging the brands of a model building sample from our tweets manually, then build predictive model(s) on the target variable(s) in Enterprise Miner. These models are using text topic variables from text topic nodes (not with user topics but normal SVD-method based ones) as predictors.

Such a diagram is shown below illustrating our process:

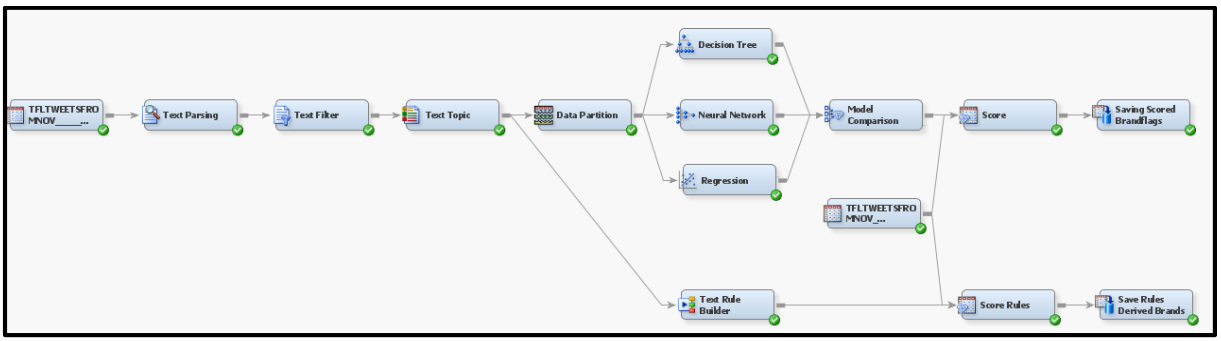


Figure 9: Diagram for predicting brand flags

The supervised learning approach shown on Figure 9 resulted in 0.61 model accuracy on the validation data set in case of the best model, which could be accepted or not, depending on accuracy needs. We also gave a try to the Text Rule Builder node. However, after looking at the tweets data set scored with the selected model, we decided to retain the rules-based approach using the user topics. Our tweets have the following breakdown per brands:

Brand	Number of tweets
Bus	364
Cab&Uber	163
Cycle	51
Rail	329
TFL General	1766
Tube	1399
Other	12

Table 2: Breakdown of brands in the tweets

In the brands breakdown in Table 2, the brand Tube means London Underground and tweets about it. TFL General is a brand category for tweets that are related to London public transport with mixed brand mentions, or about Transport for London as an organisation and a brand on its own. The category “Other” is a fallout category – it is for the tweets with the hashtag #tfl but where the user meant something other than public transport (e.g. tweeted with #tfl as a typing mistake, or about another company which is also abbreviated as TFL). The remaining brands are self-explaining.

This will be the brands assignment (brands variable) that we will use in our analyses later.

TOPIC IDENTIFICATION AND ANALYSIS

Our next analytics goal is to basically explore and identify relevant topics that tweeting is made about in our corpus.

As a first step, we used word clouds in SAS® Visual Analytics to generally explore the tweets. Visual Analytics produce word clouds with topics breakdown, where topics are produced in a less controllable way then in Text Miner. As we would like to use word clouds solely with the purpose of visualization and exploration, a different method has been followed. We used the TextFilterx_graph_table tables behind the Text Filter nodes in Text Miner via copying them from the EMWSy folders. After filtering for kept terms, we loaded the tables to LASR to produce the word clouds in Visual Analytics – but using the category values this time.

The word cloud for the brands “Bus” and “Tube” is shown below on Figure 10 and Figure 11, respectively. The size of the words is scaled with the document frequency in the actual brand, whereas the colour of the words is depending on the overall document frequency (across all brands) of the word – the bluer the word the more related it is to the actual brand.

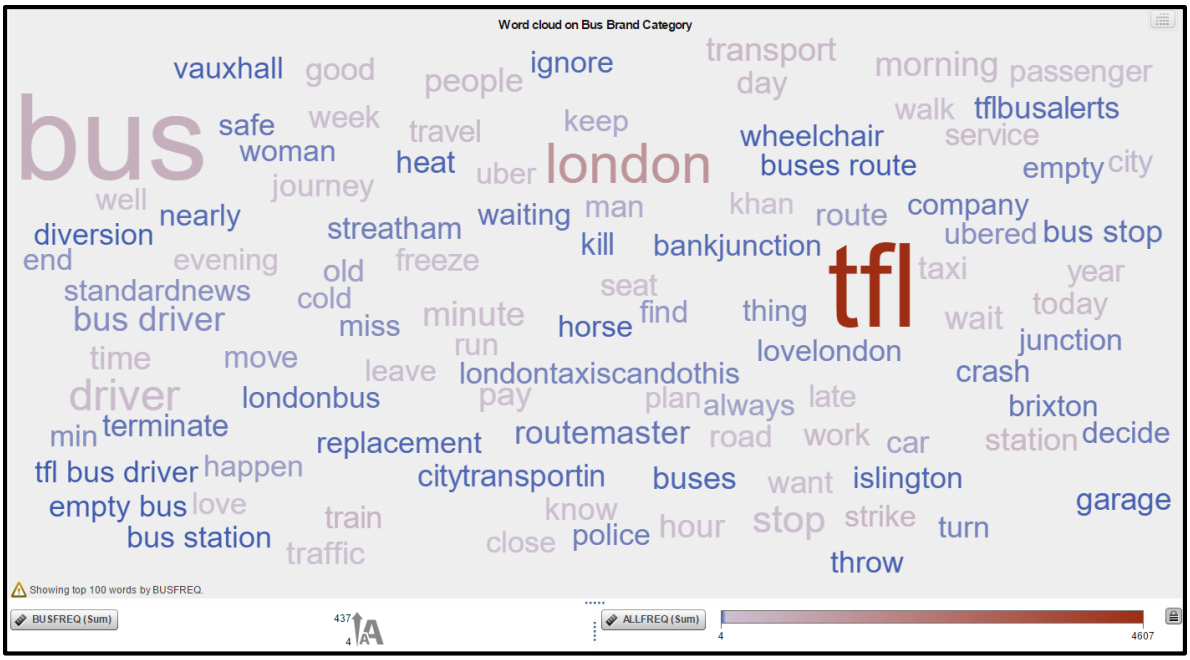


Figure 10: Word cloud on the Bus brand

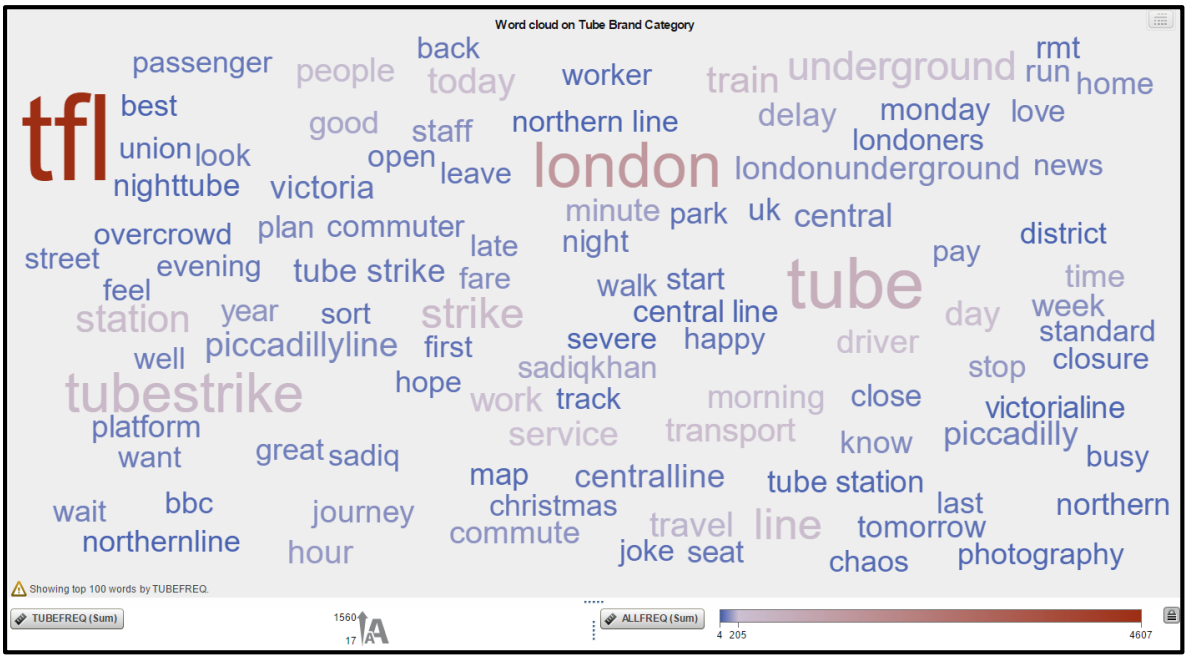


Figure 11: Word cloud on the Tube brand

In the word clouds (Figures 10 and 11) we can see that the terms used to determine the actual brands appear frequently, just as expected. This assures us in our expectation that to find topics uncorrelated with the brands in the document collection (in other words, to not just reproduce the brands with the topics again), we need to consider dropping the terms used to determine the brands from our analysis on the topics. We are using a Text Filter node again in our new diagram to drop the terms used for the rule-based brand identification from the further analysis.

Our first topic analysis showed the following results:

Topic ID	Topic	Number of Terms	# Docs
1	+strike,monday,+commuter,rmt,staff	36	646
2	+service,+good,good service,+pay,christmas	52	218
3	+transport,public,+public transport,+poster,sale	53	228
4	+day,+late,+good,+run,home	32	211
5	+morning,monday,commute,+photo,+story	39	207
6	+travel,+station,+plan,+closure,+line	33	262
7	+station,+tube station,+close,staff,victoria	47	233
8	+teacher,efl,+role,diversify,tefl	8	12
9	sadiqkhan,+freeze,+fare,+fare,sadiq	66	238
10	+minute,+wait,+stop,waiting,+platform	49	160
11	+driver,+bus driver,+tube driver,+stop,+passenger	68	214
12	+people,+seat,+stand,+walk,+good	63	168
13	today,+challenge,+station,+find,+early	48	170
14	+line,victoria,+delay,+love,monday	48	210
15	+time,first,first time,+year,+keep	50	137
16	+hour,rush,+rush hour,+journey,half	71	136
17	+work,+journey,+home,on time,+road	55	189
18	+year,commute,+happy,+seat,christmas	65	296
19	+week,+journey,+delay,next,+year	72	237
20	+work,+late,+leave,home,staff	55	192
21	standard,evening,evening standard,standardnews,evening	80	195
22	+delay,severe,+severe delay,+minor delay,minor	56	78
23	+know,christmas,+run,check,+open	66	170
24	+good,+travel, +journey,city,commuter	75	352
25	+commuter,commute,+run,+traffic,mayoroflondon	104	377

Table 3: Initial Topics in the tweets corpus

This Text Topic node result table (using the default 25 multi-term topics to keep) shows that all the topics are relevant to public transport except for one topic. Topic ID 8 has the following main descriptive terms: +teacher,efl,+role,diversify,tefl (see Table 3). This suggests an irrelevant topic and 12 documents are flagged to have this topic. Looking at the actual tweets it becomes clear that they are irrelevant to our public transport area of interest. These 12 tweets are therefore tweeted with the #tfl hashtag but as a mistake or meant to be not about Transport for London. We are going to filter out these 12 documents from our further analyses, which leaves us with 4,072 relevant tweets altogether.

The topic results above suggest that some topics might be relatively close to each other and therefore might be beneficial to merge them or to take less topics as relevant in the next run for topic identification. These merge and filter activities could be easily done in the Interactive Topic Viewer window in the Text Topic node.

In the next run for identifying the topics we tried out a couple of combinations on the weight functions to be used and the number of topics to keep in the results. Also, in some cases we considered manual topic merging. The mostly relevant result combination was the default (log for frequency weighting and entropy for term weighting) and 9 multi-term topics to be kept. This resulted in the following topics in our tweets:

Topic ID	Topic	# Docs	Topic Interpretation	Assigned Topic Label
1	+strike, today, +work, monday, rmt	724	This topic is about the tube and rail strikes which went all over the observed time period	Strike
2	+line, +service, +good, good service, +delay	444	Mainly tweets/reports and discussion about service quality and delays	Good/bad service
3	sadiqkhan, +freeze, mayoroflondon, +fare, evening standard	276	Mostly about old and new fares and the fare freeze decision by the authorities	Fare freeze
4	+day, +work, +week, +time, +work	246	Discussing usual and unusual travel times on various routes	Travel time
5	+travel, +minute, +station, +plan, +closure	409	About planned closures when customers are notified in advance	Planned closures
6	+morning, commute, +hour, monday, +people	356	General discussions about the morning rush hours	Morning rush
7	+station, +tube station, +close, staff, +station	331	Reports and opinion tweets about operative closures during the day	Operative closures
8	+transport, public, map, +year, +poster	266	About general customer information, selling of merchandising material (like printed network maps)	Merchandising
9	+driver, +minute, +wait, +time, +bus driver	434	How drivers are handling customers and different situations	Driver behaviour

Table 4: Final relevant topics in the tweets

In Table 4 above, the Topic Interpretation columns shows our interpretation and the Assigned Topic Label is a result of these.

Now we have relevant topics assigned and flagged our tweets about what brands they are referring to. We are aggregating this into one table that will have the following structure where we have our brands in the "Brands_transport" variable and all the topic flags in variables named like "SUM_of…..":

Brand_transport	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…	SUM_of_Te…
Bus	7	14	21	13	43	40	21	25	127
Cab&Uber	4	7	23	4	5	0	4	14	42
Cycle	3	2	8	2	5	4	3	7	7

Figure 12: Brand and topic summarization table

At this stage, we are about to perform correspondence analysis on the table shown on Figure 12 to project the different topics and the brands in a low-dimensional space. This is a similar method as the one described in Reference [2]. We will use PROC CORRESP to do so:

```
proc corresp data=SGFORUM.TOPICS_BRANDS_SUMMARY all outc=SGFORUM.TOPICS_BRANDS_CORDI;
ods select configplot;
var S:;
id brand_transport;
run;
```

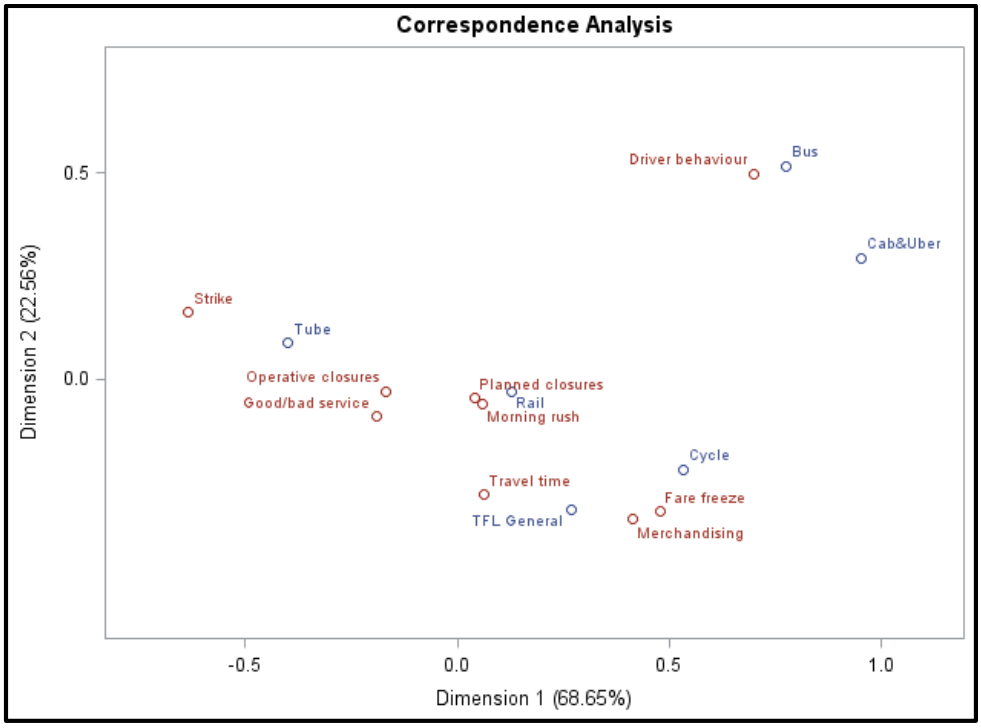


Figure 13: Correspondence analysis output

As we see on our correspondence analysis plot on Figure 13, more than 90% of the variability is explained by the two first principal components, therefore our representation of the topics and brands in a two-dimensional common space turns out to be appropriate. This correspondence analysis plot would allow us associate and group topics and brands together. However, as per our process diagram on Figure 1, we would like to perform sentiment analysis on the tweets to identify the general attitude against public transport brands and the topics presented above. We will project our brand topic maps including our sentiment analysis results later.

SENTIMENT ANALYSIS

Our process diagram identifies our next step: to perform sentiment analysis on our tweets document collection. We have multiple options to approach this.

SAS® Visual Analytics offers a sentiment analysis facility within the word cloud explorations/visualisations. It is possible to create word clouds based on variable having the role “Document Collection”. In the settings of this word cloud we can specify to identify topics and perform sentiment analysis. The result is a word cloud for each topic identified and sentiment assignments using the values: Positive, Negative and Neutral

(on the topic level, too). This is a nice and easy-to-use functionality with built in text analytics and an instant visualisation. However, we cannot extract the sentiment results in tabular form from Visual Analytics. As we would like to create custom analyses based on the sentiment assignments, we decided to follow another route.

To achieve brand flagging, we eventually used a rule-based method in Text Miner using user topics related to the brands in public transport. We are going to follow a similar method for sentiment analysis on the tweets.

As tweets are short (limited to 140 characters' length), it sounds like a reasonable approach to use sentiment dictionaries. Sentiment dictionaries are collections of positive, negative and sometimes neutral words, terms or expressions. There is not enough length in the text of tweets to have complex grammatical structures. Expressions (multi-word terms) are however beneficial to use in sentiment dictionaries as this is an easy way to account for negation (multi-word terms like "not good", "never delayed", "hardly safe"). We are using a widely available sentiment dictionary available online (reference [3]) that consists of almost 8,000 one-word term entries with a polarity assignment (negative or positive). The Enterprise Miner sample data set SAMPSIO.AFINN_SENTIMENT has some overlap with these collections of words. The sample data set also includes weight values, based on the strength of the actual terms (weights in absolute value close to 1 for very strongly positive or negative words and weights in absolute value close to 0 for very weak positive or negative words). Using this data, we create negated multi-word terms to approach capturing the polarity of tweets where negation is playing a role.

Sarcasm can also play an important role in sentiment analysis, as well as the use of emoticons (or emojis). However, looking at our corpus we decided to exclude sarcasm and emoticons from our scope as in this set of tweets these appeared to be relatively rare and therefore not playing a major role.

We are using a Text Topic node again to calculate sentiment polarity for the tweets, using user topics, and then a Save Data node afterwards to save our sentiment assignments:

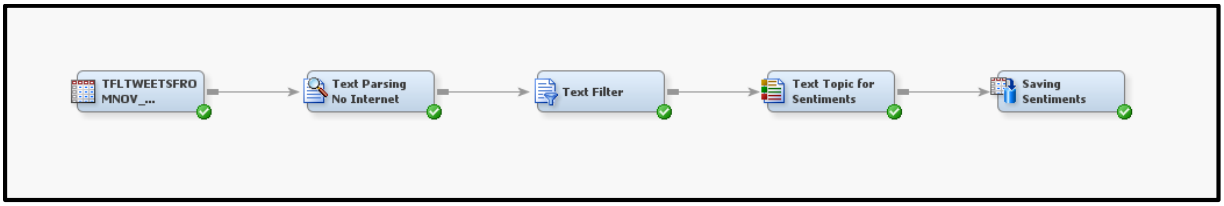


Figure 14: Diagram for sentiment analysis

Our user topics table for the sentiments altogether has 10,366 entries and is uploaded to the Text Topic node:

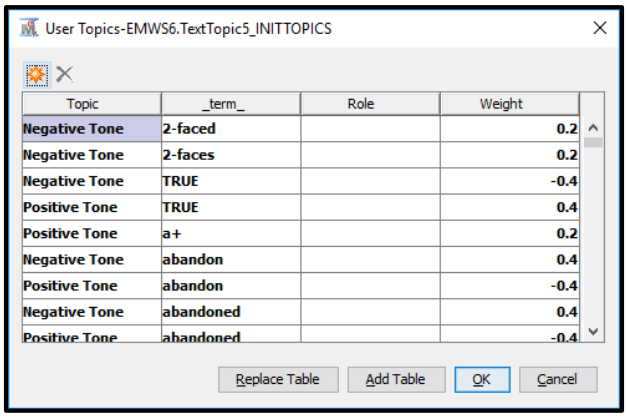


Figure 15: Sentiment dictionary table as user topics

The Text Topic node will assign negative and positive polarity values separately. If a tweet does not get assigned to neither the negative nor the positive polarity, we consider it neutral. Also, as our sentiment is not 100% balanced between positive and negative, there are relatively low number of tweets that have both negative and positive polarity values assigned. We also consider these 97 tweets as neutral. The sentiment assignments are as follows after running our Text Topic node:

Sentiment Polarity	Number of Tweets
Negative	1409
Neutral	1474
Positive	1189

Table 5: Sentiment breakdown of tweets

Our corpus is slightly skewed towards the negative sentiment – this is somehow expected as public transport is rarely a conversation topic when being satisfied with it. Rather, people can be expected to tweet more actively about public transport when perceiving negative experience. Our sentiment analysis results support this expectation.

We will analyse topics and brands altogether with sentiments. As we will visualise these on a brand topic map, we need to come up with a way to aggregate the number of negative, neutral and positive sentiment tweets per topic and brand.

In marketing theory and practice, there are a couple of sentiment aggregation indicators. Reference [4] lists example approaches. One of the common sentiment aggregation indicator is called Net sentiment. It is based on the idea of Net Promoter Score and is calculated as follows:

$$\frac{Number\ of\ positives - number\ of\ negatives}{Number\ of\ positives + number\ of\ neutrals + number\ of\ negatives}$$

The Net Sentiment value can lie between +100% (all positive in the corpus) and -100% (all negative in the corpus) for each category or category combination. As this is a symmetric indicator we will use this to visualise aggregate sentiment for our brands and topics.

CREATING MIND-MAPS

Let's use correspondence analysis results again on our brands ad topics, but this time visualise it together with the sentiments (net sentiment)! Also, it might be beneficial to show the volumes of tweets for the actual brands and topics as well.

We can easily achieve this using the `SGFORUM.TOPICS_BRANDS_CORDI` coordinates data set. To that table, the net sentiment values needs to be joined as well. We are using PROC SGPLOT to draw our brand topic sentiment map (our mind-map). As we would like to visually distinguish between brands and topics, we will use different shapes in our PROC SGPLOT:

- Topics are plotted using a BUBBLE statement. Size of the bubbles is scaled to the volumes.
- Brands are plotted as hexagons using a POLYGON statement. Size of the hexagons is also scaled to the volumes.
- Both bubbles and hexagons are coloured using the net sentiment variable. As our net sentiment distribution is skewed against the negative sentiments, the color scale would be skewed as well accordingly as we specify it in the COLORRESPONSE option for both bubbles and polygons.
- Topic labels (using mixed case for topics and uppercase for brands) are adjusted to reduce clutter.
- DRAWORDER and TRANSPARENCY options were used to be able to visualise objects close to each other.

The result plot is shown below:

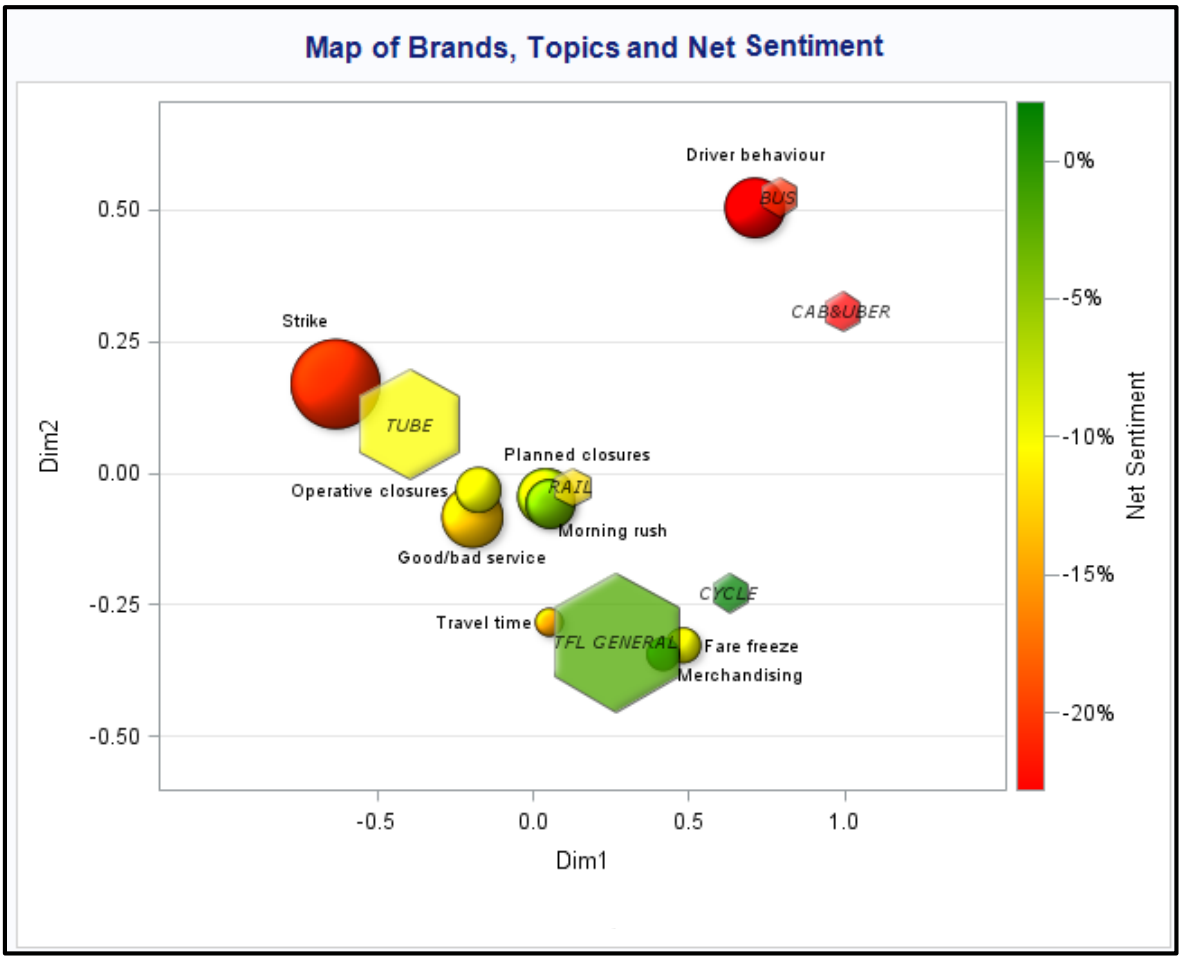


Figure 16: Brand Topic Map combined with Volumes and Sentiment

Some findings about the brand topic map in Figure 16:

- The brand Cycle seems to have a relatively positive net sentiment, cycling being popular and a healthy alternative to daily commuting. However, the weakness of the cycling infrastructure and the danger of cycling seems to have an effect as well.
- The brand TFL General is closely related to discussions about travel time and fares. Both travel time and fare topics seems to be average net sentiment (slightly negative). This can be attributed to positive experiences related to fare freezes, but also negative experience about generally high fares and long travel times. TFL General is also related to Merchandising – no surprise that people are rather relatively happy about that.
- In the middle area of the map we have a couple of topics associated with the actual quality of the services, closures and the general feelings about the morning rush. These are closely related to the Rail and the Tube brands. No surprise that all these are rather average net sentiment (slightly negative). The topic Morning rush has the best net sentiment among those in the middle area as this topic is a general discussion – whereas the other topics in the area are rather about actual service issues, hence having more negative polarity.

- The topic Strike shows very negative polarity, as expected. Strikes were happening with both Tube and Rail in the observed period, this in line with the fact that they are situated close on our map. However, the strike's negative polarity is balanced out with other topics/effects for the Tube and Rail brands.
- Although the Strike topic has very negative polarity, Driver behaviour is even more negative. The Bus and the Cab&Uber brands seems to be closely related to the topic Driver behaviour. All three have negative net sentiment compared to the average on the corpus. Negative experience about driver behaviour might be caused by politeness issues or recent reported violence cases by Uber drivers.

Based on our interpretations above, we believe specific actions could be beneficial to improve customer satisfaction and sentiment. However, they would require mainly communications actions from the public transport authority (Transport for London) instead of large scale infrastructural or organizational investments.

- As cycling is generally positively received, promoting it actively via the #tfl hashtag would likely be also received well. Combining this with some of the merchandising activities would surely mean a great combination to improve satisfaction with Transport for London. Similar actions have been carried out recently to promote walking in London (e.g. with the Walk the Tube maps) – hence adapting these to cycling might have a positive effect on sentiment.
- Strikes will always be received very negatively. However, apart from actual negotiations with the trade unions to avoid strikes, communications (tweets) about these negotiations could possibly improve sentiment at least for Transport for London (not necessarily for the trade unions). There are tweets in our document collection analyzed where Transport for London tries to distinguish itself from the trade unions to separate from the negative sentiment related to strikes – this seems to be a way of communication to continue.
- As driver behaviour is received very negatively, apart from education for bus drivers and introducing quality controls, there might be another type of action to carry out to improve sentiment. Other European public transport authorities are launching campaigns to emphasize the stamina, hard work and dedication of bus drivers via presenting actual drivers (including names, photos, work mission, dedication for safety, family etc.) to the public. These campaigns are executed on Social Media mainly and are well received, so they could positively affect the sentiment towards the bus brand generally.
- For Cabs and Uber, as the negative sentiment towards driver behaviour is affected by violence cases, two actions could be beneficial. One is to require these companies like Uber to introduce driver psychological tests which should help getting rid of violence. Also, a safety campaign is already taking place and could be suggested to continue: the campaign is promoting the usage of taxis and cabs from agencies where the drivers are already required to take examinations (like as it is the case for classical London cabs).
- Improving customer communications generally should develop the sentiment against daily commuter issues like delays and closures. This is already in practice (like live tweets containing service updates) so seems to be also a type of communication to continue in the future.

CONCLUSIONS

Social Media Data is easily collected and consumed from SAS. The Twitter Search API can be authenticated to from PROC HTTP using the OAuth 2 method. Twitter Search results can be parsed using the JSON libname engine or alternative methods to SAS data sets. These data sets can be easily used in SAS Text Analytics products such as SAS Text Miner and in SAS Visual Analytics.

Analysing social media data related to the public transport domain reveals opinion and sentiment patterns from customers, passengers or people interested in the domain. As public transport is a service which is used in daily life, people's opinions and sentiment towards it is usually negative, or in the best case, neutral. Positive sentiment towards public transport is relatively rare. However, when breaking down these social media conversations to different topics detected using SAS Text Analytics, it is clear that several areas

exist which people feel positive about. Such areas include fare discounts, introducing of new services or general customer communications. The areas people tend to think and feel negatively about in public transport include delays, service disruptions caused by technical issues or actions like staff strikes.

REFERENCES

[1] How to import Twitter tweets in SAS DATA Step using OAuth 2 authentication style

Falko Schultz, 2013.

http://blogs.sas.com/content/sascom/2013/12/12/how-to-import-twitter-tweets-in-sas-data-step-using-oauth-2-authentication-style/

[2] Text Analytics and Brand Topic Maps. Paper for SAS Global Forum 2016

Nick Evangelopoulos, 2016.

https://sasglobalforum2016.lanyonevents.com/connect/fileDownload/session/3C5FCAA6927F91492E6166DD7979591A/SASGF2016_3980_Evangelopoulos.pdf

[3] Opinion Lexicon (Sentiment Lexicon).

Bing Liu, Minqing Hu and Junsheng Cheng, 2005.

http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

[4] Social Media Metrics – SIM Score vs Net Reputation Score (NRS)

Dr Paul Marsden, 2010.

http://digitalintelligencetoday.com/social-media-metrics-sim-score-vs-net-reputation-score-nrs/

[5] SAS® Enterprise Miner and Text Miner 14.1 Help

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Tamás Bosznay
Amadeus Software
+44 (0)1993 848010
Tamas.bosznay@amadeus.co.uk
www.amadeus.co.uk