Using Social Media Data for Comparing Brand Awareness, Levels of Consumer Engagement, Public Opinion and Sentiment for Big Four Australian Banks.

Inna Kolyshkina[1], Boris Levin [2] and Grant Goldsworthy[3]

[1] School of Information Technology and Mathematical Sciences
Division of Information Technology Engineering & Environment
University of South Australia
GPO Box 2471
Adelaide SA 5001
Email: Inna.Kolyshkina@unisa.edu.au

[2] TBIS Holdings Pty Ltd
PO Box 899
North Sydney NSW 2059

[3]All Financial Services (NSW) Pty Ltd
PO Box H161
Australia Square, NSW 1215

Abstract

The growing availability and popularity of opinion-rich resources on the web led to an eruption of activity in the area of analysis of data coming from these resources. Opportunities exist to understand the extent of public engagement and sentiment toward a brand, a product or an event. In this paper, we present a case study for opinion extraction applied to the banking domain that illustrates how social media can be used to gain insight into the public opinion, sentiment and spread of social conversation related to this domain including changes that are triggered by a domain-relevant event. We applied advanced machine learning and data science techniques to the relevant social media and news data from the web to analyse the nature of public opinion in Australia toward the four major Australian banks in the context of the banks reaction to the Reserve Bank of Australia lowering the official interest rate. The resulting insights into public sentiment, reach, the topics discussed by the public and how these compared between the banks can be used proactively to inform organisational decision making.

Keywords: social media data, machine learning, random forests, generalised boosted models, text mining, R, sentiment analysis, topic modelling.

1 Introduction

Opinion mining of social media data triggered by the Web 2.0 success has been quickly developing as an important area of interest for both business and research. This paper presents a case study for public opinion extraction applied to the banking domain. The context of the study is as follows. In October 2012 the Reserve Bank of Australia lowered the official interest rate. The four major Australian banks took some time before acting then passed on rate cuts for borrowers of less than the full cut by the Reserve Bank.

The aim of the study was to discover public reactions and gain insights into the following questions:

- The volume and extent of reactions. Did people talk more about the banks as a result of the rate cut? How many people talked? How many people listened (i.e. were reached by the messages)?
- The nature of public sentiment. Did rate cuts affect consumer and media sentiment toward banks? How did consumer and media sentiment compare? Were media more or less critical of the banks than the public?
- Differentiation between the banks. How did the banks compare in terms of the number of people interested them and what people said about them?
- Public opinion groupings. What were the main topics discussed and opinions expressed about the banks? What population groups expressed them?
- Bank public relations initiatives. What levers (campaigns, community initiatives, sponsorship etc.) the banks were using to improve popularity and public sentiment? Did they work?

While the analysis was done using a number of social media sources, this paper concentrates on a methodology for data analysis using Twitter data.

This project illustrates how social media data can be harvested to gain insight that can be used to proactively manage organisational public image, brand awareness and customer satisfaction.

2 Data Extraction and Storage

2.1 Software used for data retrieval and storage

We used Java technology stack for data retrieval. The library twitter4j written in Java establishes the bridge

Copyright ©2013, Australian Computer Society, Inc. This paper appeared at the Eleventh Australasian Data Mining Conference (AusDM 2013), Canberra, 13-15 November 2013. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 146, Peter Christen, Paul Kennedy, Lin Liu, Kok-Leong Ong, Andrew Stranieri and Yanchang Zhao, Ed. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

between the client program and the data available through the Twitter public API. The Java code that is responsible for various parts of data gathering and processing is written in a manner that allows for components reuse and deployment in various standalone and hosted environments, in particular, in a web application or as part of an enterprise solution. For the storage, PostgreSQL 9.0 was chosen.

2.2 Data extraction and storage process

We collected 12 weeks' worth of social media data starting from October 3 mentioning the “big four” Australian banks (WBC, CBA, NAB and ANZ banks) and originated in Australia. The Twitter API allowed us to extract the data geographic location of the poster, user name and textual self-description. The extraction was made in bulk (as opposed to one per call), to account for Twitter's throttling restrictions, in other words we accumulated user IDs available on tweets, then issued a single call to get users data for the list of IDs.

Figure 1 gives an overall view of what happens to the social data within our solution. The steps completed in order to prepare data for analysis are outlined below.

1. Data retrieval using the API published for developers. New data was collected on a daily basis and appended to the existing data set.

2. Data collation and pre-processing. Once the data had been retrieved, we prepared it for storage. All handling of the data in Java used tweets in the form of a tweet bean created from the Twitter data seed. This bean encapsulates, on top of the standard tweet data, some elements obtained as the result of the pre-processing, such as user location and gender, followers count, etc.

3. Validation. We used data analysis in the ways described below to develop specific rules to identify and filter out entries that were considered noise and had to be removed from further processing

4. Storage. The table structure that was used to store tweets was reasonably straightforward closely resembling that of the actual tweet. The Tweeter feed provided a unique tweet ID that could also be used as the database ID. As the tweet beans already contained data enriched at the pre-processing stage, persistence logic was very simple. An object-relational mapping product Hibernate facilitated seamless integration of persistence logic into Java code.

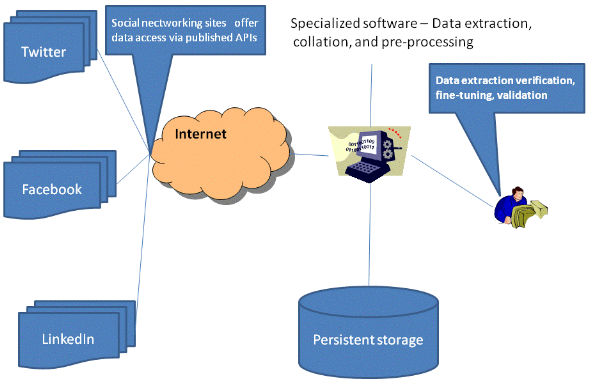
Twitter
Social nectworking sites offer data access via published APIs
Specialized software – Data extraction, collation, and pre-processing
Data extraction verification, fine-tuning, validation
Internet
Facebook
LinkedIn
Persistent storage

Figure 1: Extracting the social media data.

3 Data Analysis

3.1 Software used for data analysis

Data analysis was done using open-source software R (`www.cran.org`). We used text processing and predictive modelling techniques implemented in R packages including `tm`, `tau`, `lda`, `randomForest`, `gbm`, `earth`, `openNLP`, `wordNet`, `twitteR`, `stringr`, `plyr`, `lda` and `topicmodels`.

3.2 Text data pre-processing: cleaning and enrichment

Prior to the sentiment analysis and topic modelling we did extensive data pre-processing and enrichment of the text field.

As a first step, we filtered out the “noise data” including duplicated tweets; spam and advertising tweets (e.g. tweets sent by the banks themselves and tweets from third party organisations advertising products of a particular bank). We also excluded irrelevant tweets for example those containing abbreviations that are similar to the bank names, for example “`I cba getting up today`” — abbreviation “`cba`” is similar to “`CBA`” which stands for Commonwealth Bank of Australia.

We then performed fairly standard pre-processing of the text field (see for example Grun and Hornik (2011) and Davi et al. (2005): put all letters in a lower case, removed punctuation, stopwords, spaces, performed stemming etc. and transformed the text data field to a document-term matrix that became a basis of the input data for further analysis.

To make sure that any information describing potentially important features of a tweet had not been lost in the text pre-processing, we enriched the data with variables that described such features. For example, we added variables showing the number of characters in the tweet, count of words in the tweet, number of repeated letters (e.g. “`grrrrrreat`”), number of capital letters (e.g. “`commbank is THE BEST`”), ratio of capital letters to the number of characters in the tweet, ratio of positive to negative words, ratio of positive and negative words to the total word count, ratio of stopwords to the total word count, count of each common emoticon (e.g. smile, frown) in the tweet, number and position in a phrase of exclamation and question marks, count of each punctuation mark in a tweet (e.g. “`I don't know ...`”, “`good!!!!!!!!!`”), number and position in a phrase of emotion-expressing words (e.g. `wow`, `yay`, `lol`, `haha`, `grrrr`), swear words, negation words (e.g. `not`, `don't`, `isn't` etc), stopwords etc. as well as flags for common 3 or more-words collocations being present in the tweet (e.g. ”`reserve bank of Australia dropped interest rate`”).

Data sparsity was addressed using the approach similar to that described in Phan et al. (2008) and Feinerer et al. (2008). A useful step in data preparation which helped in sparsity reduction was combining and recoding a variety of spelling variations of frequently used words and word combinations e.g. “`thank you`” =“`thankee`” =“`thankyou`” =“`thanks`” =“`thanx`”=“`ty`” etc.

Additionally we took extra steps to further minimise the potential effect of sparsity on the validity and robustness of the findings at the later stages of the data analysis, in the process of modelling. For example in performing sentiment analysis, we applied to the data three modelling methods (generalised boosted regression, random forests and multivariate adaptive regression splines) and checked their outputs for consistency.

3.3 Sentiment scoring

The purpose of our sentiment analysis was to establish whether a tweet carried a positive and negative opinion or emotion toward the bank it mentioned. For example the tweet “lovely weather, great day, walking past Westpac building” might express positive emotion but that emotion is not directed toward the bank and so it cannot be classified as a tweet with positive sentiment for the Westpac Banking Corporation.

3.4 Preparation for sentiment analysis

As is typical for sentiment analysis (for example, see Wilson et al. (2005)), prior to the analysis we prepared lexicons, word lists and synonym lists to be used in the analysis. We created our lexicons by modifying and combining well-known lexicons available in public domain, for example, Hu and Liu’s opinion lexicon (http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar (see Liu (2010), Hu and Liu (2004)) and further enriched them to reflect the specifics of the context of our domain as well as the Australian contextual specifics.

The lexicons, word lists and synonym lists we prepared included: list of words with positive or negative polarity generally and within our context (contextual polarity may be different from the word’s prior polarity e.g. “lowered rate” is positive in the banking context but “lower” may be perceived as a negative word in general context), list of words expressing negation (e.g. “not”, “no”, “isn’t” etc); list of stopwords , list of swear words (e.g. “hell’); emoticons and words expressing them (e.g. ☺ “frown”); list of words expressing emotion e.g. (e.g. wow, yay, lol, haha, grrrr); synonym lists — general and relative to the domain (e.g. “Commbank”, “CBA” and “Commonwealth bank”) and list of important words/concepts relative to the domain (e.g. “interest rate”, “Reserve Bank” etc.)

3.5 Building a sentiment scoring model

To achieve maximum accuracy in an efficient and statistically valid way, we decided to approach assigning sentiment to a tweet as a classification problem and created a classification model for sentiment scoring.

We started from creating a rating data set similar to our data in terms of geography, domain and collection time which comprised 500 tweets. The tweets were then manually marked by two annotators as 1 standing for “positive towards the bank” (e.g. “Westpac is the best”), -1 standing for “negative towards the bank” (e.g. “Westpac is the worst”), or 0 standing for “other” (e.g. “I am now in Westpac, see u soon”). To maximise the reliability of the annotation, we marked a tweet as 1 or (-1) if the annotators agreed on the sentiment and 0 if annotators disagreed (the annotators agreed in 91% of cases).

Then, we used the annotated data to build a sentiment scoring model. To achieve this with maximal accuracy, three predictive models were applied to the annotated data: random forests, generalised boosted regression and multivariate adaptive regression splines, the target variable being the annotated sentiment score and the predictors being the elements of the document-term matrix and the added enrichment variables. The models were implemented in R using packages randomForest, gbm and earth. The models were built with cross validation on a 70% randomly selected training subset of the data and their accuracy was then tested on the remaining 30% test subset. The sentiment for the original tweet data was then calculated as the combined scores of the three models.

Apart from the derived score, another important output from the models was finding the words and phrase or sentence features that were the most important in predicting the sentiment. For example, the ratio of count of exclamation marks to the count of characters in the tweet was one of important drivers of a negative sentiment, while the number of smiles in the tweet was one of important drivers of a positive sentiment.

3.6 Establishing main common topics of public interest and the related sentiment

It was important to establish the main topics of interest to the public related to the major Australian banks and how they varied from one bank to another as well as the dynamics of public interest in the topics over time.

To establish the number and nature of the key topics of interest we applied topic modelling, the machine learning and natural language processing technique implemented using the latent Dirichlet allocation (LDA) as described in Blei and Lafferty (2009). LDA approach assumes that K latent topics are associated with a document collection, and that each document exhibits these topics with different proportions and the posterior distribution of the topics given the observed documents determines a hidden topical decomposition of the collection. The optimal number of topics K is established as the number that provides the best model fit, typically measured by a log-likelihood-based criterion (see, for example Grun and Hornik (2011)). In our case the best model fit was achieved when the number of topics was seven.

To establish the meaning of each of these topics, we reviewed the most frequent terms associated with the topic, and the results suggested that the topics were as follows: bank reaction to the interest rate drop, bank community initiatives, bank employee wages, offshoring skills, feedback on the customer service across banks, economic reports published by banks (e.g. Westpac Consumer Sentiment report) and banks’ online interfaces.

We then scored messages by whether they were related to each of the seven key topics, and based on that established sentiment per topic and compared the sentiments across the banks.

3.7 Establishing the socioeconomic characteristics of the comment-makers

To understand whether topics, reach and sentiment varied by different population groups, we needed to establish the characteristics of the posters where this was possible.

The main data field, apart from the geographical information on the poster, which was applied for derivation of such information, was the user description field. To derive user’s socioeconomic information (e.g. gender, marital status, occupation), we performed textual analysis of this field including data cleaning and pre-processing similarly to the description provided above; creating relevant dictionaries and synonym lists for this context (for example “journalist” =”journo”, “mother” = ”mom” = ”mum” =”mummy” = ”mommy”); collocation analysis to find the most common words and word combinations for this field e.g. “wife and mother”, “CEO”, “freelance journalist”, “part-time”. We then used the outputs of this pre-processing to add flags that described the features of the user description field, for example:

- Occupation information, which also may serve as a proxy for age and income, for example,. “journalist”, “student”, “CEO”, “manager”, “developer”, “part-time”, “retired”
- Gender (e.g. “mother”, “lady”)
- Family status: marital and parental status children (“husband”, “single”)
- Organisation or private poster

Gender information for private posters was then further refined by comparing the names provided in the user name field to with the list of names by gender provided by US social security website http://www.ssa.gov/OACT/babynames/.

The sparsity in user description field was managed similar to how it was done for the text field analysis described above. To further reduce sparsity, we combined occupation or family status word groups that had less than 35 posters into a broader category. For example “journalist”, ”writer”, ”editor” and ”columnist” was combined into the category called “writer_journalist’’.

In some cases user description data did not provide clear information on the occupation, age or family status of the poster. For example a poster may have had a user name like “doglover123” and user description like “Don’t walk behind me, I may not lead. Don’t walk in front of me, I may not follow. Just walk beside me and be my friend”. Preliminary textual analysis of the tweets posted by such users indicated that they were closer to private users then to organisations or news sources in terms of the use of swear words, emoticons, exclamation marks, dots and capital letters per tweet. While detailed investigation of data on this group was outside of our scope, in future it may provide more insight into its likely socioeconomic characteristics. In our analysis such posters were analysed as a separate group called “private_poster”.

3.8 Measuring Public Engagement

Key questions of interest in terms of public engagement were as follows: how many people expressed a specific opinion or sentiment and how many people were exposed to these messages.

The former was measured as the number of messages. To gain insight into the latter we used a measure called reach. In the application of statistics to advertising and media analysis, reach refers to the total number of different people (or sometimes the percentage of the target audience) who had an opportunity to see the ad. The reach of a tweet, for the purposes of this project, was defined as the maximal theoretical number of the people who were “reached” by the tweet i.e. to whom the tweet was sent in the original or retweeted form and who therefore could have read the tweet. Twitter API allows us to obtain the number of the retweets for each tweet and the number of the followers of the poster. Reach was calculated as sum of number of retweets (or slightly reworded tweets) multiplied by the number of followers of the users who retweeted. This was achieved by iterating through the list of the tweets collected since the beginning of observations, finding the number of the retweets for each tweet and the number of the followers for each of the users who post a retweet. This data was accumulated for each tweet.

4 Findings — selected examples and their statistical significance

Findings of the study illustrate the commercial insight that can be extracted from social data in terms of public opinion and sentiment dynamics and spread in reaction to events. While the complete set of the delivered findings is out of scope of this article, we present some selected findings as an example.

4.1 Statistical significance of the reported differences between the banks in terms of the established sentiment, topics and reach

Comparison of the banks in terms of the reach, sentiment and coverage for each of the identified seven key topics of interest involved formal testing of the statistical significance of the differences between the banks with the null hypothesis of no difference existing in each case. These null hypotheses were tested by the analysis of variance approach.

The results indicated the presence of significant differences between the four banks in terms of the expressed sentiment ($p>0.05$) and the reach ($p<0.01$). Some of the topics such as offshoring skills and low bank employee salaries were significantly more expressed in relation to certain banks ($p<0.01$). For other topics, e.g. the topic related to economic reports produced by banks, there was no significant differences across the banks ($p>0.2$).

All the differences mentioned below are statistically significant unless stated otherwise.

4.2 Event-driven dynamics of public opinion. Influence of RBA’s rate drop on public interest and public opinion toward banks

The RBA rate drop drew public attention to the big four banks. The total reach of bank-related messages significantly increased by 40% in the weeks following interest rate drop ($p<0.01$).

The big four banks’ not matching fully the RBA interest rate drop caused a significant ($p<0.05$) fall in overall sentiment (5%) towards the banks in the two weeks after the event with the sentiment starting to improve in the third week

4.3 Public engagement findings summary. Comparison of banks by number of messages and reach

Bank 1 seemed to be the most popular bank to be discussed on Twitter with 31.0% all banking-related messages mentioning the bank. It was closely followed by Bank 2 (29.0%). However reach was higher for Bank 2 (33.2%) than for Bank 1 (26.3%) which suggests that Bank 2’s social media strategy was more effective than Bank 1’s. Bank 4 was third in public engagement with 23.8% messages mentioning the bank; however reach was low (9.0%) suggesting opportunities to improve the bank's social media strategy. Bank 3 seemed to have had the lowest level of social media engagement among the four banks with only 5.6% media messages relating to the bank and 8.6% of bank-related media reach.

4.4 Sentiment analysis findings summary

Sponsorship and community initiatives-related messages improved sentiment toward banks while the most negative sentiment was related to banks' interest rate drop not matching the RBA level, Bank 1 not treating employees fairly, Bank 2 off-shoring skills and Bank 3's poor online interface as well as consumer complaints across the banks.

We summarised public sentiment by posters' gender, occupation, marital status etc derived from the data as described above. The categories with the significantly different from the average sentiment were as follows. Among the private posters, the consumer group that had markedly higher sentiment than the rest were executives (managers, CEOs, CTOs etc.). This group expressed 9.4% higher sentiment than the average, positively commented on rates and bank stocks and did not make customer service complaints. The consumer groups that had markedly lower sentiment than the rest were firstly IT professionals who commented on banks' online interfaces inadequacy and were disappointed by Bank 2's IT skills offshoring and secondly those describing themselves as married or having children who expressed disappointment by the banks failing to meet the RBA level of rate cuts, were sympathetic to the low-paid banking employees and made a number of customer service complaints. Overall, private posters expressed significantly ($p<0.05$) lower sentiment than organisations and media sources.

4.5 Main seven topics of interest across banks and associated sentiment

The established main topics of interest are listed here starting from the topic with the highest reach and ending with that of the lowest reach. The topics included banks' reaction to the interest rate drop by RBA (mostly negative sentiment), banks community initiatives (neutral/positive sentiment), bank employee wages (negative sentiment), banks offshoring skills (negative sentiment), feedback on customer service across banks (mostly negative sentiment), economic reports published by banks (neutral sentiment) and banks' online interface (mostly negative sentiment).

4.6 Customer service feedback and complaints summary

Insights from the analysis of the spontaneous consumer feedback can be directly used by banks in improving customer experience and maintaining their market share.

The level of customer service provided by the banks was among the key topics of public interest. Consumer comments were providing feedback on banks' customer service online, via ATM and in the branches. Of this, up to 80% was negative. The positive feedback focused on helpfulness of staff in the branches (particularly for Bank 1). The complaints focused on banks' online interface deficiencies (particularly that of Bank 4), customer service in branches (most negative feeling being created by Bank 2), email spamming (particularly Bank 1), ATM issues and customer fees (no significant difference across banks).

5 Conclusion

In this article we present an example of a process for extracting social media data about Australian banks, analysing it and arriving at insights about the topics being discussed, the spread of discussion, the sentiment expressed and the dynamics of the above. Our approach capitalises on modern-day technology and machine learning advances on the one hand, and on the power of the open source software movement, on the other. Having applied the methodology to a real situation that followed the rates drop, we were able to discover mainstream societal responses, as well as to capture finer nuances of how public reacted to that specific event and to certain aspects of the banking industry as a whole. The methodology described can be easily adapted and applied in other studies. The techniques can be extended to deal with larger volumes of information and with more complex analysis requirements. Overall, the business community would benefit from reliable feedback on their marketing effort, industry initiatives, promotion campaigns, etc. that can be obtained from social media using these techniques.

References

Blei, D. and Lafferty, J. (2009), *Text Mining: Classification, Clustering, and Applications*, Chapman and Hall/CRC Press, chapter Topic Models.

Davi, A., Haughton, D., Nasr, N., Shah, G., Skaletsky, M. and Spack, R. (2005), 'A review of two textmining packages: SAS TextMining and WordStat', *The American Statistician* **59**(1), 89–103.

Feinerer, I., Hornik, K. and Meyer, D. (2008), 'Text mining infrastructure in R.', *Journal of Statistical Software* **25**(5), 1–54.

Grun, B. and Hornik, K. (2011), 'Topicmodels: An R Package for Fitting Topic Models', *Journal of Statistical Software* **40**(13), 1–30.

Hu, M. and Liu, B. (2004), 'Mining and summarizing customer reviews.', *KDD-2004* .

Liu, B. (2010), *Sentiment analysis and subjectivity. Handbook of Natural Language Processing*, second edn.

Phan, X., Nguyen, L. and Horiguchi, S. (2008), Learning to classify short and sparse text & web with hidden topics from large-scale data collections, *in* 'Proceedings of the 17th International World Wide Web Conference (WWW 2008)', Beijing, China, pp. 91–100.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005), Recognizing contextual polarity in phrase-level sentiment analysis, *in* 'Proceedings of HLT-EMNLP-2005'.