Examining the accuracy of sentiment analysis by brand monitoring companies

Author: Qin Li
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
q.li-5@student.utwente.nl

ABSTRACT. Sentiment analysis is prevalent in the recent years as a business evaluation and prediction approach. An increasingly number of companies is participating in the sentiment analysis market to monitor sentiment related to a certain company; the service is called brand monitoring service. However, the accuracy of sentiment analysis has not been evaluated. This paper addresses six criteria of accurate sentiment analysis and selects four companies to examine the degrees of theoretical applications in the reality.

Supervisors: Fons Wijnhoven and Ton Spil

Keywords
Sentiment analysis, opinion analysis, brand monitoring, accuracy measurment, polarity classification,

5th IBA Bachelor Thesis Conference, July 2nd, 2015, Enschede, The Netherlands.

1. INTRODUCTION

Nowadays, user interactions are active on the internet, where buyers share the experience of a certain product or service and potential customers make their purchasing decisions based on the reviews by experienced users. The comments online contain subjective sentiments and opinions (Bhadane, Dalal, & Doshi, 2015; Pang & Lee, 2008). Sentiments influence the sales volumes of the product or service to some extent. Companies view that the detection of sentiment about certain product or service lead to competitive advantage, through which they catch the opportunities and weaken the threats (Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015; Wijnhoven & Bloemen, 2014). To capture the public opinion about a brand, sentiment analysis assists to extract sentiments from tens of thousands of expressions written online and classifies them into different value groups. This requires much work done in constructing algorithms and the system. In this sense, brand monitoring services provided by specialized companies to help other companies monitor the sentiments about their products or services, as well as competitors'.

Sentiment analysis is a natural language processing (NLP) application, detects author's attitudes, emotions and opinions from different areas of texts in real-time by applying artificial intelligence (Bagheri, Saraee, & de Jong, 2013; Kranjc et al., 2015; Pang & Lee, 2008). There are two main approaches for conducting sentiment analysis – machine learning and lexicon-based approach (Medhat, Hassan, & Korashy, 2014; Serrano-Guerrero et al., 2015). Machine learning is based on the selection and extraction of an appropriate set of features used to detect sentiments (Medhat et al., 2014). While lexicon-based approach relies on the pre-defined lexicon dictionaries and/or datasets (Kang & Park, 2014). The widely applied technique for sentiment analysis is polarity classification by which the sentiment is divided into three values: positive, neutral and negative (Pang & Lee, 2008).

However, some researchers argue that the sentiment words placed in each value cannot reflect the real opinion of the author (Kranjc et al., 2015; Pang & Lee, 2008). In addition, sentences contain a sophisticated semantic structure. A single word cannot determine the overall sentiment orientation of a sentence (Pang, Lee, & Vaithyanathan, 2002). Another problem related to sentiment analysis is that people may not use standard words or grammar to express their opinions, which increases the difficulty in analysing the sentiments. Besides, the comments under different domains represent different sentiments. An example of sentiment distinction in different domains shown by Pang and Lee (2008). "Go read the book" is a negative expression in the movie review domain, while it represents a positive opinion in the book review domain. The contradictory result shown by brand monitoring companies may lead the client company overconfident its performance. In this context, the accuracy of sentiment analysis is a crucial criterion for firms to develop a partnership with the brand monitoring companies. Therefore, this paper will study "How do brand monitoring companies cope with the accuracy of sentiment analysis?"

The structure of this paper is that Section 2 explains different problems related to sentiment analysis. Section 3 describes sentiment analysis in detail and the measurements of sentiment accuracy developed by existing literature. Section 4 demonstrates the methods used in this study. Next, section 5 evaluates the most important four brand monitoring companies and gathers data from the selected companies. Section 6 establishes the results. Finally, section 7 draws a conclusion and states limitations and future research.

2. PROBLEM DISCRIPTION

The first aspect of the problem is that the sentiment contains various levels, which cannot simply be measured by three values (Pang & Lee, 2008). In many cases, placing words in polarity classification does not reflect the appropriate value of the emotion (Pang & Lee, 2008). Since the value of a particular word is influenced by the context, instead of determining by the word itself (Kranjc et al., 2015). Besides, there are many words convey the same sentiment, but the degree of the emotion is various. When people describe a positive feeling, they can use, for instance, not bad, good and wonderful. The real degree of above positive feeling varies, and sometimes "not bad" is put in neutral or even negative side if the algorithm only recognizes "bad" (Pang & Lee, 2008). In addition, no positive or negative word shown in the sentence also conveys sentiment orientation, for instance, 'I missed the bus, because of the out of date timetable'.

The second problem is that people use ungrammatical sentences (i.e. lack of punctuation) and non-lexicon expressions (i.e. gr8, :)) online. In this case, the sentiments may not be correctly identified, because of the underdeveloped algorithm. Falling in the identification of sentiments leads to an inappropriate classification of the sentence, even results in contradictory meaning.

Thirdly, words express different meaning under different domains. Although the lexicon databases, such as SentiWordNet, provide the value of a word, it does not apply in a specific domain (Park, Lee, & Moon, 2015).

3. LITERATURE REVIEW

3.1 Sentiment Analysis

Sentiment analysis is also known as opinion mining or sentiment mining. Sentiment analysis focuses on identifying the positive or negative value of opinions or expressions are written in natural language (Jang, Sim, Lee, & Kwon, 2013).

Since companies have different focuses and purposes of sentiment analysis, they choose either machine learning or lexicon-based approach. Machine learning is an active learning process by combining existing data sets with newly discovered unlabelled expressions (Kranjc et al., 2015). The lexicon-based approach depends on dictionaries and datasets, where the polarity or property of a single word has been defined (Park et al., 2015). The basic activity of sentiment analysis is to extract opinion words and phrases from texts. However, sentences which do not contain any opinion word or phrase are also conveying sentiment. In machine learning, researchers have found a way to deal with this problem. Medhat et al. (2014) use the so called - features selection to demonstrate a appropriate and accurate sentiment analysis. There are five features 1) *terms presence and frequency* evaluate and calculate the words occur in mentions; 2) *parts of speech (POS)* finds important adjectives; 3) *opinion words and phrases* are elements determine the meaning of the sentence; 4) *negations* include negative words, which may change the sentiment orientation of any sentence. In some studies, in order to acquire the overall sentiment for a sentence or document, the polarity classification is presented in numerical order, range from -1 to 1 to evaluate each word within the text and then sum up the total scores got within certain sentence or document (Kang & Park, 2014; Maks & Vossen, 2012; Pang & Lee, 2008).

3.2 Accuracy

Knowing the accuracy of the sentiment analysis is beneficial for both monitoring companies and their (potential) clients. On one hand, brand monitoring companies can alter or improve the algorithm for precise analytical results. On the other hand, the higher the accuracy of sentiment analysis is, the more customers choose to work with.

Serrano-Guerrero et al. (2015) develop an algorithm to measure the accuracy of the result $\frac{\#\mathrm{hit}}{\#total\_reviews}$, where #hit means the number of reviews that were correctly classified. #total_reviews represents the total number of reviews tested. In this way, acquiring original sources of all the posts is the precondition.

Based on the research of Saif, He, Fernandez, and Alani , they conclude that the use of supervised learning method, a subset of machine learning, increase the accuracy of sentiment analysis due to assistance from the large number of sentiment detection websites and continuous learning for sentiment classifier, such as Naïve Bayes.

Jang et al. (2013) develop Personality-value-attitude (PVA) model to take customer's profile into consideration to enhance the accuracy of sentiment analysis, to know their behaviour and needs. They believe that the polarity method has its own limitation that the calculation method is not universally applied to all the posts online.

4. METHODS

In order to study the accuracy of brand monitoring services, a collection of related literature is introduced first. In search of the literature, the following keywords are used, [sentiment analysis or opinion mining or sentiment mining and brand monitoring and accuracy] in Scopus and ScienceDirect. This paper is focused on sentimental expressions in English, thus, the literature which research on the language other than English is not being reviewed. In addition, this paper will focus on both machine learning and lexicon-based approach and will establish a mutilcriteria approcah for evaluating the accuracy of sentiment analysis, by combining the criteria have developed from previous literature. The first criterion is polarity scoring. Opinion word is not the only element to determine the sentiment of a text or document. Pang and Lee (2008) find that when combine words with no obvious emotions into a sentence, it can show a clear sentiment orientation. In this sense, every word in the post should be analysed and give a numerical value to calculate the overall sentiment. The second criterion is the multiple data source. Petz et al. (2013) believe that one source contains bias, because the authors can only represent a small group of people who have the same or similar characteristics. Consider domain specification as the third criterion because many articles have proven that words and expressions under different domain affect their meaning, leading to different sentiment orientation (Pang & Lee, 2008). Non-lexicon expressions are widely used on the internet. So the capability to detect the non-lexicon is another criterion to evaluate sentimental accuracy. In linguistics, double negation represents affirmation. However, current sentiment analysis tools do not correctly detect the double negation expression, for example, not bad, classified to negative sentiment in most cases (Medhat et al., 2014; Pang & Lee, 2008), so negation is selected as the fifth criterion. The last criterion is customer profile, which do not directly affect accuracy of sentiment analysis, but give contribution to construct a strong cognition on behaviours behind the words and customers' needs in order to provide a more accurate sentiment analysis program (Jang et al., 2013).

Therefore, this paper addresses six criteria of sentiment accuracy from literature and is summarized as follows:

Polarity scoring – this measures the overall sentiment of a sentence or text by combining all the sentiments within the sentence or text.

Multiple data sources – the capability of the sentiment analysis extracts data from different media.

Domain specification – expressions represent different meanings in different domains.

Non-lexicon evaluation – the non-word expressions can be detected and classified to appropriate category.

Negation – the negative words should be detected and transferred to the correct polarity.

Customer profile – tracking the customer's behaviours or feedbacks helps to understand their true feelings.

By typing [brand monitoring service], [brand monitoring] and [sentiment analysis] in Google to search for companies provide the services.

The information about sentiment analysis is provided by brand monitoring companies is not sufficient on their websites, therefore, a questionnaire is sent out to acquire related information (see Appendix).

In order to avoid bias, this paper will focus on the comments, reviews and expressions of Samsung – a multinational conglomerate company, developing fast in electronic industry, in April, 2015 for all possible sources written in English online. "gr8" and ":)" are used to test whether the brand monitoring tools can identify the non-lexicon expressions and "not bad" is used for negation test.

5. ANALYSIS

5.1 Brand Monitoring Companies

This paper examines four brand monitoring companies, which provide sentiment analysis as part of their business. The information is presented in Table 1 and Table 2.

5.1.1 Socialmention

Socialmention is a lexicon-based analysis tool, provides brand or keyword detection from various social media. It shows the sentiment ratio (generally positive to generally negative), top authors and sources. Socialmention allows people to track back the original posts.

Socialmention detects 146 mentions about Samsung written in April, and there are 24 positive mentions and 1 negative mention, the rest are neutral. By calculating the sentimental orientation, the number of positive posts is divided by the total number of negative posts.

Socialmention is not able to analyse the symbol - :), no result shown while it is capable to evaluate "gr8".

When analyse "not bad", the result demonstrates a high negative value. 111 out of 157 posts are classified to negative.

5.1.2 Mention

Mention is a company deals with brand/keyword extraction and sentiment analysis. It allows to process text from many resources, like Blog, Twitter, Facebook. Sentiment classification is based on predefined lexicon datasets.

Mention presents 182 results of Samsung in April. There are 80 positive mentions while 6 negative mentions. Mention uses the scale to show the number of posts in polarity value. 2,266 mentions are collected by Mention for "gr8".

There is no result presented for ":)" and "not bad".

5.1.3 Topsy

Topsy is an analytical company, collecting tweets, according to their polarity. Regarding sentiments, it divides sentiments into two values: positive and negative.

Topsy collects 2,410,800 tweets that mentioned Samsung in April. The overall sentiment analysed by Topsy is 65% positive, which means the number of positive mentions is a little higher than which of negative mentions.

In general, "gr8" expresses positive feelings. Although there is a small percentage of tweets of "gr8" shows a negative opinion, the overall sentiment is high, which is 91% positive.

Based on the sentiment score of "not bad" provides by Topsy, we can conclude that Topsy is not able to analyse Negation. The result shows an extreme negative sentiment score, 18 points.

5.1.4 Trackur

Trackur is a social media monitoring company, adopting machine learning approach. Trackur uses its own analytical tools to monitor sentiment and reputation, providing monitoring services for individuals, small and large companies and agencies. It tracks words from multiple social media, such as Facebook, Instagram, and Twitter.

Trackur collects 2,939 posts about Samsung, where 594 mentions are positive and 151 mentions regard as negative. Trackur is the only company which presents detection of symbol ":)", 2,988 posts are found, 672 positives and 266 negatives. 368 mentions of "gr8" are detected, where 106 positive and 19 negative posts. Trackur shows a negative sentiment tendency for "not bad", in which 529 out of 786 mentions are negative and 117 posts are positive.

When search for ":)" and "gr8", Trackur shows the image below. It adds the new words to their database instead of saying "no result".

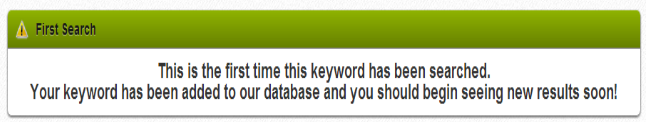


Figure 1: First Search Notification

Companies	Criteria					
	Polarity scoring	Multiple data source	Domain specification	Non-lexicon evaluation	Negation	Customer profile
Social mention	X	O	X	-	X	O
Mention	X	O	X	-	X	O
Topsy	X	X	X	-	X	O
Trackur	-	O	X	O	-	O

Note. O: Applicable; X: Not Applicable; -: Not Completely Applicable

Table 1: Companies Summary

Keywords	Socialmention	Mention	Topsy	Trackur
Samsung	146 (P: 24; N:1)	182 (P:80 ; N:6)	2,410,800 (P: 65%)	2939 (P:594 ; N:151)
:)				2988 (P:672 ; N:266)
gr8	108 (P: 11; N:0)	2,266 (P:594 ; N:23)	338,395 (P: 91%)	368 (P:106 ; N:19)
Not bad	157 (P:22; N:111)		1,394,296 (P: 18%)	786 (P:117 ; N:529)

Note. P: Positive; N: Negative

Table 2: Detailed Search Results for Each Company

5.2 Comparison

The information about the method that determines the sentiment of words or sentence are absent among all the companies. Although Trackur shows that it analyses sentiment of keywords in sentence level and then applies the sentiment to the whole document, to increase the possibility of obtaining the real sentiment of the author, the in depth explanation does not present.

Topsy classifies sentiment into two values – positive and negative, neglecting neutral mentions, while, the other three use three-value classification. In the three-value classification, we have seen that the number of neutral posts occupies a certain quantity.

All the brand monitoring companies can detect keywords from multiple public sources, except Topsy, which is worked on Twitter posts. Topsy uses single sources to collect mentions, but it gathers the largest number of mentions during the given period.

All the companies in this research do not provide domain search.

As for the non-lexicon expression, "gr8" is an informal expression of "great". "Great" has a lower tendency to be defined as a negative word, which of the probability is 0.125[1]. Three companies, Mention, Topsy and Trackur, do collect a small percentage of negative posts, 0.01, 0.19 and 0.05 respectively. Socialmetion does not gather negative mentions.

By testing the ability of detection and placement of negative words, "not bad" is used. To our knowledge, "not bad" does not represent negative meaning, though "no" and "bad" are negative words. However, all the three companies which are able to detect "not bad" shows an overall negative value.

Providing the name or page link of authors is an extra assistance for people who want to know what the author really thinks about certain products or services. Mention, Topsy and Trackur display the top related or active user for each keyword (see Appendix). Besides, by creating a customer profile, people can find out the customer's behaviour, in order to filter real customers and spam accounts.

5.3 Error Analysis

Two-value polarity classification puts the neutral posts into either positive or negative, but Topsy does not provide the criteria for classifying the neutral mentions to positive or negative, resulting in a relatively higher extreme (positive or negative) value than which of the other three.

The analysis tools cannot correctly detect ":)". Besides, on some websites, there are emotion insertion charts, which are not belonging to the target texts, but the analysis tools include the chart. Moreover, the symbol ":(" incorrectly is recognised as ":)" by the tools and gives a negative value for the post (see Figure 2). Trackur shows that first search notification and will add these into their database. The first time search can be problematic since the keywords are not classified to a certain value, until staff updates the programme.

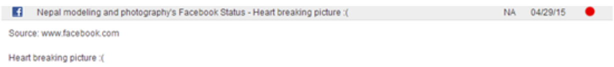


Figure 2: Incorrect Detection of :) from Trackur

There is an example from Trackur, which shows negative value for "not bad", but by checking the original posts, we see that the author do not mean any negative sentiment in that context.

[1] http://sentiwordnet.isti.cnr.it/search.php?q=great

Figure 3: Negative sentiment placing for "not bad" from Lovell Soccer

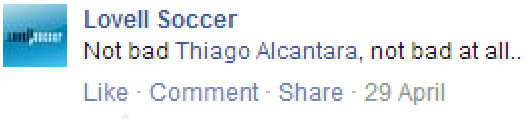


Figure 4: Original post of Lovell Soccer

6. RESULTS

By comparing the number of keywords detected during April, Topsy collects the largest number for each keyword, even though Topsy is a Twitter focused analysis company. Socialmention provides the least amount of posts. Consequently, the quantity of keyword detection is not determined by the number of sources the company has.

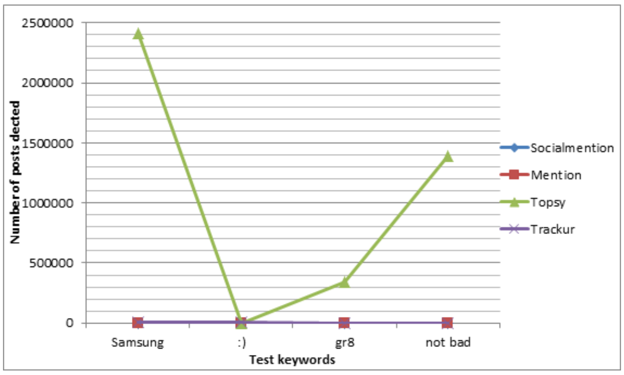


Figure 5: The Number of Mentions Detected by Each Company

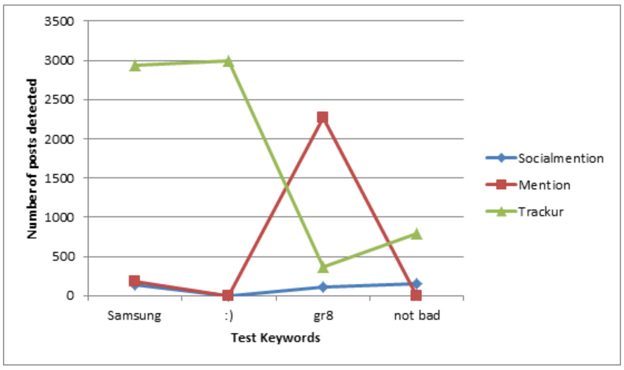


Figure 6: Detailed Number of Mentions Collected by Socialmention, Mention and Trackur

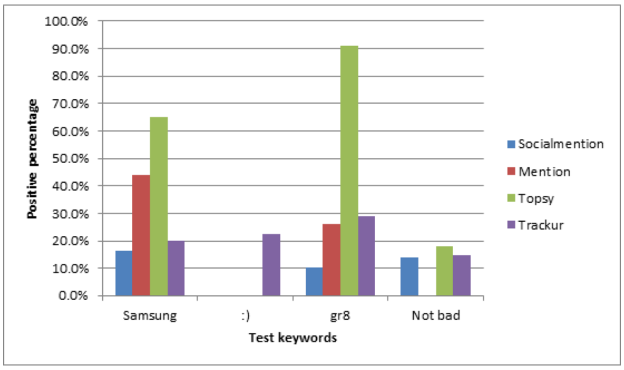


Figure 7: Positive Percentage of Each Keyword Presented by Four Companies

Topsy gives the highest positive percentage of "Samsung" and "gr8" and "not bad". Socialmention and Trackur display that authors did not show positive attitudes to "Samsung" in April, accounted for 16.4% and 20% respectively, but Topsy gained 65% positive sentiment. Topsy also gave an extremely high positive sentiment of "gr8", which is 91%.

	Socialmention	Mention	Topsy	Trackur
"not bad" mentions	157 N: 70.70%		1,394,296 N: 82%	593 N:62.56%

Table 3: Negative percentage of "not bad" shown by each company

All the companies provide a more than 50% negative orientation to "not bad". By tracking 20 negative posts from each company, the result shows that all of them indicate "not bad" as a negation opinion, which is contradictory to our common knowledge.

Companies	Summary
Socialmention	✧ Three-value polarity classification ✧ No polarity scoring ✧ Multiple sources ✧ No domain specification ✧ Informal word detection ✧ No symbol detection ✧ Problematic negation classification ✧ Low quantity of mentions detected for all the keywords ✧ Individual sentiment result and link provided ✧ Top active authors presented but no grouping
Mention	✧ Three-value polarity classification ✧ No polarity scoring ✧ Single word analysis ✧ Multiple sources ✧ No domain specification ✧ Informal word detection ✧ No symbol detection ✧ No active authors presented
Topsy	✧ Two-value polarity classification ✧ No polarity scoring ✧ Single source - Twitter ✧ No domain specification ✧ Informal word detection ✧ No symbol detection ✧ Problematic negation classification ✧ High quantity of mentions detected for all the keywords ✧ Top active authors presented but no grouping
Trackur	✧ Three-value polarity classification (Sentence-level) ✧ No polarity scoring ✧ Multiple sources ✧ No domain specification ✧ Symbol and informal word detection ✧ Problematic negation classification ✧ Individual sentiment result and link provided ✧ Top active authors presented but no grouping

Table 4: Results

7. CONCLUSION

This study demonstrates six important characteristics related to the accuracy of sentiment analysis in the literature. A case study has been conducted to evaluate the characteristics applied in the real world by selecting four brand monitoring companies, which provide free trails. It has presented that no companies include all the criteria or keep it as a business secret. The most important finding is that symbols cannot be detected. Three out of four companies are unable to collect posts contained ":)", and Trackur is incorrectly detected the symbol. Besides, "not bad" is regarded as a highly negative phrase among all four tools. Problems and errors are found during the experiment. Mention and Trackur claim that the accuracy of their sentiment analysis result is more than 70%. However, incapacity of detecting and analysing symbols and double negation phrases is highly influence the accuracy. The accuracy information provided by the companies is not as high as what they claim. On the other hand, the literature findings do not widely be adopted in the real world. Companies should put more efforts in constructing the basic algorithm and interacting with outside experts to improve the accuracy of sentiment analysis.

The major limitation in this paper is that analysing the sentiment given by the companies exists bias. Each company provides the number of positive and negative posts, but the real true number of classifications cannot be evaluated. Not all the companies present the link of original mention. Another limitation is the posts accessed by brand monitoring tools are public while there are a number of accounts and pages prevent to be seen by strangers.

Future studies can focus on what algorithm uses to detect sentiment or keyword on different social media channels. And how to measure the accuracy of multilingual sentiment analysis can be researched.

8. ACKNOWLEDGEMENT

This research was conducted as a bachelor thesis for the International Business Administration programme at University of Twente. I wish to express my gratitude to my supervisor, Mr. Fons Wijnhoven, for his admirable guidance and insight into this research. I would also like to thank Jérôme Garnier (Mention) and Andy Beal (Trackur), without the support from them, this research would not have been presented.

9. APPENDIX

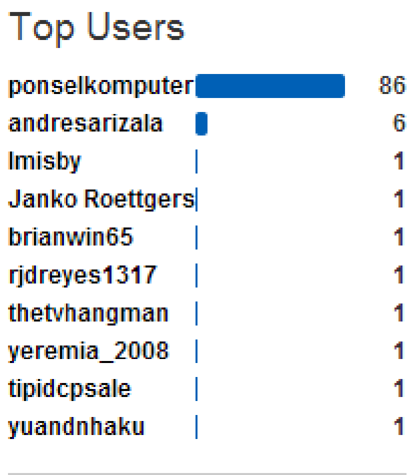
Appendix 1: Questionnaire

Questionnaire

1. What kinds of SA approaches do you apply in your service?
   a. Lexicon-based method
   b. Machine learning
   c. Dictionary-based method
   d. Corpus-based method
   e. Supervised learning
   f. Unsupervised learning
   g. Other -
2. What are the sources of sentiment analysis?
   a. Facebook
   b. Twitter
   c. Video
   d. Images
   e. Blogs
   f. Forums
   g. Other -
3. How do you collect the data from media?
4. How often does your SA collect data?
5. What kind of analysis platform do you use, existing analysis platform (eg. SentiWordNet) or self-established platform?
6. When the non-lexical structures (eg. ☺, :-), gr8) occur, how does the algorithm analyse?
7. What detection level does your SA achieves/operates?
   a. Sentence level
   b. Entity(topic) level
   c. Domain level

How do you evaluate the accuracy of your SA results?

Appendix 2: Top Users Identification from Socialmention



Appendix 3: Top Users Identification from Topsy

10. REFERENCES

Bagheri, A., Saraee, M., & de Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems, 52*(0), 201-213. doi: http://dx.doi.org/10.1016/j.knosys.2013.08.011

Bhadane, C., Dalal, H., & Doshi, H. (2015). Sentiment Analysis: Measuring Opinions. *Procedia Computer Science, 45*(0), 808-814. doi: http://dx.doi.org/10.1016/j.procs.2015.03.159

Jang, H.-J., Sim, J., Lee, Y., & Kwon, O. (2013). Deep sentiment analysis: Mining the causality between personality-value-attitude for analyzing business ads in social media. *Expert Systems with Applications, 40*(18), 7492-7503. doi: http://dx.doi.org/10.1016/j.eswa.2013.06.069

Kang, D., & Park, Y. (2014). Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. *Expert Systems with Applications, 41*(4, Part 1), 1041-1050. doi: http://dx.doi.org/10.1016/j.eswa.2013.07.101

Kranjc, J., Smailović, J., Podpečan, V., Grčar, M., Žnidaršič, M., & Lavrač, N. (2015). Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Information Processing & Management, 51*(2), 187-203. doi: http://dx.doi.org/10.1016/j.ipm.2014.04.001

Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems, 53*(4), 680-688. doi: http://dx.doi.org/10.1016/j.dss.2012.05.025

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113. doi: http://dx.doi.org/10.1016/j.asej.2014.04.011

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr., 2*(1-2), 1-135. doi: 10.1561/1500000011

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up?: sentiment classification using machine learning techniques.* Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

Park, S., Lee, W., & Moon, I.-C. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters, 56*(0), 38-44. doi: http://dx.doi.org/10.1016/j.patrec.2015.01.004

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V., & Holzinger, A. (2013). Opinion mining on the web 2.0–characteristics of user generated content and their impacts *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 35-46): Springer Berlin Heidelberg.

Saif, H., He, Y., Fernandez, M., & Alani, H. Contextual semantics for sentiment analysis of Twitter. *Information Processing & Management*(0). doi: http://dx.doi.org/10.1016/j.ipm.2015.01.005

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences, 311*(0), 18-38. doi: http://dx.doi.org/10.1016/j.ins.2015.03.040

Wijnhoven, F., & Bloemen, O. (2014). External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews? *Decision Support Systems, 59*(0), 262-273. doi: http://dx.doi.org/10.1016/j.dss.2013.12.005