Influential Opinion in South African Tweets: Should business continuously monitor customer sentiment and its potential influence in Twitter?

A research report presented to
The Graduate School of Business
University of Cape Town

In partial fulfilment of the requirements for the
Masters of Business Administration Degree

by
Vicente Da Silva
December 2012

Supervised by: Walter Baets
Contents

1. Determining the value of monitoring sentiment and its potential to influence ................................................. 4

2. A review of the current research .......................................................................................................................... 6
   2.1. Twitter, Sentiment, Influence; what is it and what does it mean to companies? ............................................. 6
       2.1.1. Twitter ............................................................................................................................................... 6
       2.1.2. Sentiment .......................................................................................................................................... 7
       2.1.3. Influence .......................................................................................................................................... 8
       2.1.4. Business applicability of this kind of analyses ..................................................................................... 8
   2.2. Sentiment and Influence Analysis, the technical details .................................................................................. 10
       2.2.1. Sentiment Analysis ............................................................................................................................ 10
           2.2.1.1. Machine Learning Algorithms .................................................................................................. 10
           2.2.1.2. Lexicon Based Approaches ..................................................................................................... 11
           2.2.1.3. Machine Learning vs. Lexicon ................................................................................................. 12
       2.2.2. Influence Analysis ............................................................................................................................. 13
   2.3. Research review wrap-up ................................................................................................................................ 13

3. Influential Opinion analysis ................................................................................................................................... 14
   3.1. Acquiring the Twitter dataset ......................................................................................................................... 14
   3.2. Identifying companies or brands with a high degree of interaction ............................................................. 15
       3.2.1. Word frequency analysis ............................................................................................................... 15
       3.2.2. Business / brand keywords ........................................................................................................... 15
   3.3. Determining and analysing sentiment ........................................................................................................ 17
       3.3.1. Shoprite and Computicket ............................................................................................................... 17
       3.3.2. Shoprite ........................................................................................................................................... 18
       3.3.3. Spar .................................................................................................................................................. 19
       3.3.4. Pick n' Pay ...................................................................................................................................... 19
   3.4. Determining influencers ............................................................................................................................... 19
   3.5. Analysis insights ........................................................................................................................................... 20

4. Conclusion ............................................................................................................................................................ 21

5. Future research ....................................................................................................................................................... 22
   5.1. Influential Opinion Model ............................................................................................................................ 23
       5.1.1. Variable 1: Sentiment deviation from normal ...................................................................................... 23
       5.1.2. Variable 2: Re-Tweet score .............................................................................................................. 23
       5.1.3. Calculating the Influential Opinion score ......................................................................................... 23

6. References ............................................................................................................................................................. 24
Tables

Table 1: Analysis of username @garethcliff ................................................................. 15
Table 2: Top 25 companies and brands ranked by number of relevant Tweets found ........ 16
Table 3: Bottom 25 companies and brands ranked by number of relevant Tweets found ...... 16
Table 4: Search keywords used .................................................................................. 16
Table 5: Total Tweets for companies selected ................................................................ 17
Table 6: Total Tweets for companies selected excluding sentiment = 0 and first and last days Tweets .... 17
Table 7: Total Tweets found on the 19th ........................................................................ 18
Table 8: Average sentiment for the period ..................................................................... 19
Table 9: User Tweet details for selected Tweets ............................................................. 20
Table 10: Weights used in influence of user sentiment calculation .................................. 23
Table 11: Weights used for re-Tweet score ................................................................... 23

Figures

Figure 1: Twitter Poll Results ...................................................................................... 15
Figure 2: Total Tweets found for Shoprite and Computicket .......................................... 17
Figure 3: Sentiment Sum for Shoprite and Computicket ............................................... 17
Figure 4: Tweets distribution over the sample period for each company ......................... 18
Figure 5: Tweets sentiment sum per day over the sample period for each company ........ 18
Influential Opinion in South African Tweets: Should business continuously monitor customer sentiment and its potential influence in Twitter?

Vicente Da Silva

Abstract: This research demonstrates that the degree of customer-brand/business interaction depends on the type of business being analysed. For some South African companies there is sufficient interaction in Twitter to warrant adapting their public relations processes to allow for continuously monitoring Twitter. This kind of monitoring is especially relevant during significant events a company experiences. In order to answer the posed question managers must, firstly, determine if this technology is applicable to their business in the South African context, secondly, they must have a basic technical understanding of concepts like Sentiment and Influence Analysis, thirdly, they must be able to determine how opinionated and influential their customers are in social media, fourthly, they must be able to determine who the influencers in their business social ecosystem are, fifthly, they must be able to analyse their competitors social ecosystem in order to compare their own to their competitors and lastly, they must have an understanding of how to adapt their business processes to cater for this additional source of customer data. This research analysed the South African Twitter stream for a period of 30 days. From the analyses 3 South African companies in the same industry were selected to do a sentiment and influence analysis on. This was done in order to get a better understanding of the nuances and quality of real world Twitter data, originating from South African, for similar companies, in order to answer the question posed. Before determining which companies to use, the entire dataset of 21 million Tweets was searched using a list of publically listed South African company names and a list of popular South African brand names. This was done in order to find a sample of companies that have a high degree of customer-brand/business interaction in Twitter. The results were analysed quantitatively and 3 companies, that had a sufficient degree of interaction, were chosen to do further analysis on. For each of these companies a qualitative analysis of the discovered Tweets was done in order to determine the nature of the sentiment and influence being expressed. Based on the insights discovered from this analysis, advice is offered to help companies understand what the implications on their business are, of incorporating this kind of analysis into their existing public relations processes. Advice is offered to companies wishing to outsource this kind of data analysis, in terms of what sort of questions they should be asking vendors of this kind of data analysis. Doing this research also laid the groundwork for the creation of an experimental model that could be used to quantitatively measure the Influential Opinion of a set of Tweets.

1. Determining the value of monitoring sentiment and its potential to influence

Many businesses have been slow to take advantage of the data being generated by the use of new communications paradigms, like social media, that their customers are using. Managers are getting caught up in the hype being created about the benefits of using social media.

A review of the literature in terms of what value there is to business in monitoring social media streams was done. There is strong evidence to suggest that this kind of data analysis can be beneficial in improving customer relations, however most of this research has been done in countries other than South Africa. The purpose of this research is to discover insights relating to the current South African Twitter ecosystem, by measuring not only the Sentiment being expressed in it but by analysing the influence potential of a sample of users in Twitter. An
analysis on 3 South African companies was done in order to discover these insights which helped to
determine if there is value to South African businesses in continuously monitoring Twitter streams.

For managers who need to decide on how, if at all, to incorporate these new paradigms into their business there are six important issues to consider.

Firstly, are they familiar enough with the various techniques employed to do Sentiment and Influence Analysis in order to ask the right questions when confronted by a social media analytics vendor claiming to have the most sophisticated technology? By better understanding the technology, managers can turn this knowledge into a competitive advantage in that they will be able to choose the kind of analytics and metrics that maximise value in the context of the social media profile of their business.

Secondly, do they understand the degree to which their customers are using social media? By understanding the degree and quality of their customer’s interaction with social media they will be able to make sound decisions as to what investment to make in monitoring social media streams.

Thirdly, the use of social media in countries like the USA, where the Sentiment Analysis technology has primarily been developed, far exceeds the usage in South Africa. In countries like the USA the volume of social media usage is great enough to make Sentiment Analysis a credible metric to be used in Public Relations (PR) activities. If South African businesses are considering using Sentiment Analysis based on metrics derived in other countries, then is there is sufficient usage in South Africa of social media, like Twitter, that would enable South African managers to derive real value from insights gained from Sentiment and Influence Analysis?

Fourthly, if the degree of social media engagement is determined to be significant for a particular businesses customers, then how can this new source of customer data be incorporated into the businesses existing processes?

Fifthly, in addition to this managers should monitor the Influential Opinion of their competitors in order to compare their competitors result to their own.

Lastly, managers need to understand what makes some interaction more influential than others. They need to understand how to identify influencers in their social media ecosystem.

The initial research did not include influence analysis. The question was then asked; is determining only the sentiment of relevant Tweets a good enough indicator of the potential impact social media commentary can have on a business or brand? Not all Tweeters are alike and some will be more influential than others. The initial research was then augmented with research on social media influence. An experimental model was created to measure Influential Opinion. This model uses an adjusted sentiment score based on the posters past sentiment compared to the sentiment of the found Tweet and the number of re-Tweets the posters Tweet has had. The Influential Opinion model uses these variables to determine an Influential Opinion score for a Tweet or a set of Tweets that have been found to be relevant to a business.

In order to provide insights into the issues discussed above it was necessary to develop a deeper understanding of the following concepts; understanding the characteristics of Twitter data, understanding what research has been done on Sentiment and Influence Analysis, what role can Sentiment and Influence Analysis have in the public relations strategy of a business.

The literature review includes a review of the current techniques used to do Sentiment and
Influence Analysis. This was done because it was felt that managers who choose to outsource analysis services rather than do the analysis internally should understand the fundamentals of the technical concepts in order to determine if the analysis services they are planning to use have been developed using appropriate techniques. The technical detail included in this paper goes only as far as is necessary for a manager to be able to understand the nuances and differences of the techniques.

The analysis was done in the context of a single social media channel, Twitter. Tweets from users originating in South Africa were used. In order to find a sample of companies to do a deeper analysis on, the entire Twitter dataset was searched using South African company and brand keywords. From these results three South African companies were chosen. A quantitative Sentiment Analysis was done followed by a qualitative analysis of the potential influence of Tweets relevant to the businesses. This results of this analysis were then used to answer the question posed by this research paper. This combination of Sentiment and Influence Analysis is referred to as Influential Opinion in this paper.

Some advice is offered to help managers understand what the implications are on their business of incorporating this kind of analysis into their existing public relations processes as well as advice to companies wishing to outsource this kind of analysis to data analysis vendors.

The primary assumptions made in this research are that Twitter data collected has been reverse-geo-coded by Twitter accurately and that someone who Tweets positively or negatively about a business is a customer of that business.

2. A review of the current research

2.1. Twitter, Sentiment, Influence; what is it and what does it mean to companies?

2.1.1. Twitter

Twitter\(^1\) has more than 100 million users and receives more than 250 million Tweets per day (Harrison, 2012). This is an enormous amount of data being generated daily. If this data can be effectively analysed, there is a great potential to discover new insights into a vast array of topics including how people interact socially and what people’s opinions are as well as developing models to predict future events. Future flu outbreaks have been predicted by using Twitter data and searching for related keywords. The results showed a 95% correlation to statistics collected by the U.S. Centres for Disease Control and Prevention (Savage, 2011). (Oh & Liu Sheng, 2011) analysed a stock micro-blog and were able to show a relationship between the sentiments being expressed and stock price movements. An advantage of using Twitter data for market research purposes is that it is relatively easy to collect information quickly and cheaply. Market research thinking changes from an asking approach to one of listening.

The research by (Savage, 2011) identified the following three problems with Twitter data. First, Twitter data is noisy but this noise is reduced as the amount of data used in the analysis increases. Second, Tweets can become viral. Tweeting about a sudden outbreak of flu can lead to more mentions of that keyword than actual cases of people having flu. Lastly, different demographics of Tweeters use Twitter more than others but researchers can lessen this bias by applying corrections to the data. (Patino, Pitta, & Quinones, 2012) identified the following five

\(^1\) [http://www.Twitter.com](http://www.Twitter.com)
problems with using social media data like Twitter data for market research purposes. Firstly, there is a belief that it will take over from traditional methods, however traditional methods are still needed to understand customers, the competition and product awareness. Secondly, unlike research panels, it is difficult to determine who is writing the commentary. It is difficult to know if the same people are posting on multiple sites. Thirdly, there is an issue of data quality and external validity, especially if the same people are posting multiple times then the data is not generalizable. Coders have to interpret the writers meaning as they do not have the option of questioning the writer. Fourthly, each target market you wish to survey use social media differently. Millennials are highly engaged users, while baby-boomers are less engaged. Lastly, ethics is becoming an important issue. Businesses are sometimes unwilling to share information on what data they are collecting as well as how they collect it.

Being able to utilise the large amount of Twitter data to determine the sentiments of social media commentators has been gaining traction in linguistics research. Initial usage of this technology is in being able to determine commentator’s sentiments regarding business and products.

2.1.2. Sentiment

Sentiment Analysis involves the use of natural language processing. This involves computational algorithms making sense of human generated commentary; however current techniques struggle to recognise ambivalent, sarcastic, or ambiguous texts. This issue is to some extent overcome by analysing large amounts of data, as in the case of Twitter feeds. In large amounts of data these types of texts become less of a factor (Harrison, 2012). Sentiment Analysis has yet to achieve high levels of accuracy. Today’s tools interpret sentiment correctly 75 to 80% of the time (Jacobson, 2009). The tools for extracting public opinion from the text of social media are crude and remain in their infancy. The extent to which these methods could supplement or replace traditional public opinion polls is unknown (Association for Computing Machinery, 2010). Even though Sentiment Analysis has yet to achieve high levels of accuracy practitioners interviewed by (Jacobson, 2009) felt that the tools will mature in time. They also felt that Sentiment Analysis is one of many tools that PR professionals should consider in determining the state of their social media relationships. This view was reiterated by (Patino, Pitta, & Quinones, 2012).

(Gayo-Avello, 2011) identified the three following issues with using current Sentiment Analysis tools. Firstly, there exists a big-data fallacy. Even though the amount of data collected can be large it is important to know if the collections are statistically representative samples of the overall population. Secondly, one must be wary of demographic bias. Without knowing user age the results may exhibit bias because the main users of social media tend to be relatively young. Researchers must try and correct for this bias. Thirdly, one must be wary of simplistic Sentiment Analysis. Some application may be able to achieve reasonable results; however researchers should avoid noisy instruments. This seems to contradict the view by (Harrison, 2012) who felt that a large amount of data can tend to overcome any shortcomings in the current tools and that using more sophisticated tools with large amounts of data tends to improve accuracy significantly. Thirdly, silence speaks volumes. Not knowing the degree to which non-responses sentiment plays in the research being done can potentially invalidate the research. Lastly, past positive results do not guarantee generalization. The degree to which social media has penetrated the population being studied is another important factor to consider. If not considered results may be questionable and possibly incorrect.

Together with Sentiment analysis another important topic to consider is Influence Analysis.
2.1.3. Influence

(Bakshy, Hofman, Winter, & Watts, 2011) found that the variables most influential in creating viral messages (cascades) are users who have been influential in the past and users who have a large number of followers. They hypothesise that it is more cost effective, for a wide range of plausible assumptions, for marketers to target individuals with average or less than average influence as opposed to individuals with a large influence. They also claim that most word-of-mouth dissemination of information happens via many small cascades initiated by ordinary individuals.

(Bakshy, Hofman, Winter, & Watts, 2011) summarise early works on influence research. They state that the effect of extensive word-of-mouth information dissemination has been found to affect public opinion (Katz & Lazarsfeld, 1955), innovation adoption (Rogers E. M., 1995), size of market share (Bass, 2004) and brand awareness (Keller & Berry, 2003).

(Romero, Galuba, Asur, & Huberman, 2011) claim that knowing who the influencers in one’s network are is necessary in order to implement viral marketing and to set the topics in social media that are discussed most. (Kardara, Papadakis, Papaoikonomou, Tserpes, & Varvarigou, 2012) reiterate this view.

In a contrasting view (Watts & Dodds, 2007) found that one of the reasons for messages going viral are not due to influential but due to a critical mass of easily influenced individuals.

2.1.4. Business applicability of this kind of analyses

In the research by (Pang & Lee, 2008), in which they provide a detailed description of the techniques used for sentiment analysis, they describe the possible use for such analysis by business and government intelligence. Business can use such analysis for reputation management, public relations and possibly even predicting sales. The current state of Sentiment Analysis tools has not prevented business from trying to use the tools to find a competitive advantage. (Bollena, Mao, & Zeng, 2011) showed that there was a correlation between peoples moods and movements in the stock market. They were able to predict the daily up and down changes in the closing price of the DJIA with an accuracy of 87.6%. This kind of research provides impetus for businesses looking for a competitive edge. According to (Simmons, Mukhopadhyay, Conlon, & Yang, 2011) 35% of Wall Street firms are now using Sentiment Analysis to analyse unstructured online news reports, editorials, and business websites. Some firms even trade on the results of the analysis. (Sonnier, McAlister, & Rutz, 2011) demonstrated that sentiment in online communications has an effect on daily sales performance. They found the effect of positive and negative comments greater than neutral comments and that positive comments have a greater effect than negative comments. They also found that the effect of comments dissipated after about a week. They claim their results are the first to show the linkage between sentiment and sales. The idea that Sentiment Analysis has the ability to improve a business’s bottom line is shared by (Wright, 2009) who also feels that Sentiment Analysis should be able to identify specific issues that a business’s customer have and help them to respond with the correct marketing and PR strategies.

In terms of the types of sentiment being expressed and how sentiment is propagated, (Simmons, Mukhopadhyay, Conlon, & Yang, 2011) found that negative messages are shared at a much higher rate than positive messages (17 vs. 11). Negative messages can be spread quickly and organisations need to respond effectively to this negativity in order to prevent negative impact on the organisations economics. Interestingly, the view that negative messages can spread quickly was confirmed by research conducted by (Naveed, Gottron, Kunegis, & Che Alhadi, 2011) where it was found that sentiment plays an important role.
in Re-Tweeting and that negative sentiment Tweets tend to be Re-Tweeted more often. (Hansen, Colleoni, Etter, & Arvidsson, 2011) found that negative sentiment news Tweets are more likely to be Re-Tweeted than negative sentiment non-news Tweets, while positive non-news Tweets are more likely to go viral. Interestingly, in the article by (Arndt, 2012) Bluefin, CEO Deb Roy discovered that a :-) emoticon will be used, on average by a commenter who is 10 years older than a commenter who used the emoticon :).

The current landscape of Sentiment Analysis techniques seems to be one of many Sentiment Analysis techniques that cater for determining the sentiment focusing on products products (Pang, Lee, & Vaithyanathan, 2002) (Blitzer, Dredze, & Pereira, 2007) (Baccianella, Esuli, & Fabrizio, 2010) but not as many used for informal communications in social media (Pak & Paroubek, 2010). Informal communication in social media tends to be much shorter and contains numerous, non-standard spelling. The exchanges tend to be short and serve to make an initial contact or to keep in touch from time to time. (Thelwall & Wilkinson, 2010)

A study by (Rogers K. , 2012) of how BP reacted via Twitter to the Gulf Oil Spill crisis in 2010 is a good example of how not to use social media to respond to a crisis. It was found that 80% of the Tweets BP made were image restoration Tweets while 20% were simply informational. Before the crisis the BP Twitter account was fairly inactive. BP used social media in a reactive way. BP should have shown a greater acceptance of responsibility and shown empathy toward the victims. The consensus seems to be that BP did not handle the crisis effectively and seemed careless in their responses. They used Twitter to disseminate information rather than to engage with individual concerned citizens. Twitter is a 2 way communication channel, BP did not seem to understand this. The PR department at BP did not seem to understand or react to the actual concerns of the people trying to engage with the business via Twitter. While this study is limited to only one social media channel (Twitter) it does demonstrate why business needs to have an integrated communication strategy across all social media channels and that the way in which business reacts has huge implications for the brand value of that business. This view is reiterated by (Shahim, 2011).

(Owyang, 2011) showed that 76% of social media crises faced by businesses could have been averted or diminished if the businesses had been better prepared. Even advanced businesses are still deficient; in that they struggle with fragmented technology, non-standardised measurement frameworks and that the customer data gleaned from social media have not found their way into the product roadmap or into customer support systems.

(Hanna, Rohm, & Crittenden, 2011) describes a social media ecosystem involving digital and traditional elements making it easier to understanding and conceptualise online social media. They describe 5 lessons learned in implementing a social media strategy: visualise the ecosystem, identify and track key performance indicators, understand clearly what it is you want to tell customers, social media does not require elaborate budgets and be unique by customising user engagement (Owyang & Li, 2011) speak of the concept of Social Business Maturity and based on it the business should budget appropriately for their social business investments. They identified three maturity levels, Novice, Intermediate and Advanced. Novice businesses should focus on getting their internal teams in order, intermediate businesses should increase the level of customer facing initiatives and advanced businesses must integrate social business throughout the business. (Woodcock & Green, 2010) speak of social CRM, making use of social media to maximise the value from Customer Relationship Management investments across the entire value chain of the business. Business must be wary of three issues; these are
organisational readiness, over-hype and over-expectation and project management failings. The author also makes the point that at the end of the day CRM is about people and relationships and the media channels are simply tools to enhance these two concepts. (Kaplan & Haenlein, 2011) discuss how social media can add value along all three stages of the marketing process; firstly, in the pre-purchase stage you should know what your customers are saying, secondly, in the purchase stage you should disseminate brand reinforcing messages and finally, in the post-purchase stage, you should try to improve customer service and complaint management. They then discuss what they believe are the three rules of micro-blogging, Relevance – focus on messages that are relevant to your target audience, Respect – show respect to your audience and finally Return – any investment in social media must be able to show a return. In the practitioner article by (Singh, 2011) the following guidelines were identified for business wanting to improve the bottom line by harvesting social media data: constantly monitor and measure brand/product perception – not just mentions, compare brand/product perception with competitors’ counterparts, measure the effectiveness of marketing programs, engage with customers and/or prospective customers and create better product designs by understanding customers’ likes and dislikes for various features.

(Dowling & Weeks, 2011) make the point that, Salience and Sentiment Analysis is really only the first step to developing a holistic view of social media analysis. They add Theme and Contradiction Analysis and Problem and Solution Analysis. Theme and Contradiction analysis involves understanding the mental model the media uses to describe the organisation and in understanding this, the business can frame their responses in a way that fits this mental model. Problem and Solution Analysis involves understanding how the media has created a problem and then devise solutions to overcome this problem. While the article is geared toward media in relation to the press, the principles discussed here are relevant and easily transferred to the world of social media.

An issue for business to consider is that of Ethics. In the study by (Tan , et al., 2011) the assumption was made that if users follow each other on Twitter then they agree with each other’s points of view / opinions or that they have a personal relationship. The study does not try and establish what percentages of users do in fact connect for these reasons vs. simply connecting to discover the point of view of others. Businesses need to be transparent in how they collect the data, how it is stored and how they analyse the data, however in a highly competitive environment where businesses may risk exposing propriety intellectual capital by being transparent, businesses will find themselves in a dilemma.

Finally the issue of credibility of commentators on a social media channel like Twitter should also be considered by business. (Castillo, Mendoza, & Poblete, 2011) propose a technique using a supervised classifier to measure the credibility of a Tweet. Business should consider not only analysing the Sentiment within social media by also the credibility of the source of the Tweet. This is important because these types of social media channels can be used to disseminate misinformation of false rumours, which can spread quickly.

2.2. Sentiment and Influence Analysis, the technical details
Sentiment Analysis is predominantly done using either machine learning algorithms or a lexicon based approach. In some cases a mixture of the two techniques are used.

2.2.1. Sentiment Analysis
2.2.1.1. Machine Learning Algorithms
There are three machine learning algorithms that are used predominantly when doing machine
learning in the context of text classification. They are Naive Bayes, Maximum Entropy (MaxEnt) and Support Vector Machines (SVM). Naive Bayes has been shown to be optimum for certain problem classes with highly dependent features; however MaxEnt and SVM algorithms often give better results. MaxEnt sometimes outperforms Naive Bayes at standard text classification (Pang, Lee, & Vaithyanathan, 2002). Naive Bayes, MaxEnt and SVM are supervised algorithms known as more generally as classifiers. Supervised algorithms mean they need to be trained on pre-tagged data and tested against a different set of pre-tagged data. When this pre-tagged data is manually tagged the algorithm is used in a fully supervised mode. When the data is automatically extracted and then manually analysed and refined the algorithm is used in a semi-supervised mode. Unsupervised mode would be when the training data is created automatically without any manual intervention. Once the classifiers achieve an acceptable level of accuracy then they can be used on un-tagged data (Bird, Klein, & Loper, 2009). When training a classifier, the most important decision is the selection of features to be used to train the classifier (Mishne, 2005). Overall the Naive Bayes algorithm seems to be the better machine learning algorithm as opposed to SVM for short texts like Tweets. Researchers were able to binary classify Tweets with 74.85% accuracy (Bermingham & Smeaton, 2010).

In the research by (Hansen, Colleoni, Etter, & Arvidsson, 2011) a naive Bayes classifier was used and trained on the Brown corpus using the NLTK toolkit. They were able to achieve an accuracy of 84% ± 1% on the test data. The authors do not specify how they filtered the Tweets, or whether their filter took the nuances of Twitter data into account as described in (Thelwall & Wilkinson, 2010).

(Pak & Paroubek, 2010) uses a Naïve Bayes classifier with N-gram and POS (parts of speech) tags as features. Bi-grams were found to have the best accuracy. The authors do recognise that there is a conflict between their research and other research in the fact that in their research bi-grams result in greater accuracy. In contrast to this, work by (Pang, Lee, & Vaithyanathan, 2002) showed unigram to have greater accuracy, however the work by (Pang, Lee, & Vaithyanathan, 2002) was based on movie reviews rather than micro-blogging data. The work of (Dave, Lawrence, & Pennock, 2003) is mentioned where it was found that bigrams and trigrams give better accuracy on product-review polarity classification. This is evidence of the need for a classifier to be trained in the particular domain it will be operating in, to achieve accurate classification. (Liu, Li, & Guo, 2012) showed that by adding the data derived from emoticons to that of manually labelled Tweets (fully supervised) improved the accuracy of the classification algorithm. They were able to obtain accuracy between 75% - 76%.

2.2.1.2. Lexicon Based Approaches
(Taboada, Brooke, Tofiloski, Voll, & Stede, 2011) define a lexicon based approach as using dictionaries of words annotated with the words semantic orientation (polarity) to determine the semantic orientation of a text. A lexicon based approach to Sentiment Analysis is a unsupervised approach, in that there is a predefined set of features that the analysis will be done with as opposed to a machine learning approach where the classifier first needs to be trained and tested, preferably using data from the same domain.

(Turney, 2002) used a lexicon based approach to Sentiment Analysis by estimating the semantic orientation of each phrase (containing adjectives or adverbs) in a text and then classifying the text using the average semantic orientation of the phrases in the text. In the research it was shown that accuracies between 66% and 84% could be achieved depending on the domain. This seems to imply that a lexicon based approach is equally sensitive to the domain it is being used in as machine learning algorithms are.
Another lexicon based approach is provided by (Cui, Zhang, Liu, & Ma, 2011) in which they used emotional tokens (irregular forms of words, punctuation and emoticons) to determine Sentiment Analysis and thereby create a multilingual Sentiment Analysis algorithm. Comparing this technique to semantic lexicons and some “state of the art” Twitter Sentiment Analysis web service they receive effective results.

(Gayo-Avello, 2011) used the Twitter Search API to collect Tweets and chose not to use a classifier based approach to analysing the sentiment. The techniques used were one based on the semantic-orientation method (lexicon based approach), one based on mention counts and two based on polarity lexicons. In this research the author readily admits that one should be wary of simplistic methods for analysing data. In this research it was perhaps the simplicity of the lexicons chosen to do the analysis rather than the lexicon based approach that yielded below optimum results.

(Bollen, Mao, & Zeng, 2011) used a mood lexicon called OpinionFinder (subjectivity lexicon) and the Google-Profile of Mood States (GPOMS) to determine mood of commentators. They were able to correlate movements in the closing value of the DJIA with an accuracy of 87%.

(Baldoni, Baroglio, Patti, & Rena, 2012) propose a method of enhancing an ontology, in this case the OntoEmotion ontology, an ontology of emotional words, which in turn can be used to do Sentiment Analysis. While this research is still highly experimental it holds a lot of promise in that the accuracy of Sentiment Analysis can be greatly enhanced by the use of ontologies where concepts have relationships to other concepts (a graph) and where ambiguities can be resolved.

2.2.1.3. Machine Learning vs. Lexicon

(Thelwall, Buckley, Paltoglou, & Cai, 2010) developed an algorithm utilising a lexicon based approach call SentiStrength2. This algorithm detects not only sentiment but the degree of the sentiment. The algorithm takes into account the grammar and spelling styles of social media. The algorithm is able to detect positive sentiment strength with 60.6% accuracy and negative sentiment strength with 72.8% accuracy. The positive sentiment detection is considered better than a wide range of general machine learning approaches while the negative Sentiment Analysis accuracy is considered similar to other machine learning techniques. (Thelwall, Buckley, & Paltoglou, 2011) developed a technique using a lexicon based approach which required no training, however the lexicon can be tuned to a particular domain being investigated. (Thelwall, Buckley, & Paltoglou, 2012) followed up their previous article in which they derived the SentiStrength algorithm. The improved SentiStrength algorithm performs better than a baseline approach for all data sets in both supervised and unsupervised cases. The algorithm was tested on six different social web domains. The software is freely available for academic use.

(Paltoglou & Thelwall, 2011) propose a similar technique to (Thelwall, Buckley, Paltoglou, & Cai, 2010) in that they use emotional lexicons and incorporate modifiers to determine the emotional intensity and strength of textual utterances. Their results show that an unsupervised algorithm outperforms other machine learning solutions in the majority of cases. This software is not freely available. This is one of the first papers to directly compare a lexicon based approach to existing dominant machine learning approaches with multiple test collections. It also demonstrates that using an effective lexicon is comparable to machine learning techniques.

(Baccianella, Esuli, & Fabrizio, 2010) show the improvement in their Sentiment Analysis toolkit called SentiWordNet 3.0 a lexical resource,

\(^{2}\) http://sentistrength.wlv.ac.uk/
however they compare the results to SentiWordNet 1.0 rather than against other machine learning algorithms. On the other hand they claim their algorithm is licensed to over 300 research groups, demonstrating the confidence academia has in a lexicon based approach.

### 2.2.2. Influence Analysis

(Kwak, Lee, Park, & Moon, 2010) compared three variables in order to establish if there is a relationship between the variables and therefore determine key indicators of influence. The variables were number of followers, page-rank and number of re-Tweets. It was found that there is a gap between number of followers and the popularity of ones Tweets (re-Tweets). They also found a positive relationship between a user’s number of followers and a user’s number of Tweets made.

(Cha, Haddadi, Benevenuto, & Gummadi, 2010) used three variables to measure influence number of followers, number of re-Tweets and number of mentions. It was found that the number of followers had little effect on the popularity of a user, as measured by the users re-Tweet count. They also discovered that most influential users can hold influence over a variety of topic and that influence is gained by concerted efforts like limiting ones Tweets to a specific topic.

(Weng, Lim, Jiang, & He, 2010) observed that in their Twitter dataset 72.4% of the users follow more than 80% of their followers and 80.5% of users have 80% of users they are following, follow them back. This phenomenon of “reciprocity” in Twitter in terms of following ones followers can perhaps explain why the number of follower’s measure is not a good indicator of a user’s influence. They developed a model called TwitterRank. Their TwitterRank algorithm builds on the PageRank algorithm but includes the topical similarity between users into account. The assumptions is that users interested in certain topics will have more influence on followers interested in the same topics as opposed to users interested in other topics. They compared their algorithm to three others and find that it performed best.

Overall the research has shown that the number of followers attribute of a Twitter user is not an accurate indicator of the influence that user has.

### 2.3. Research review wrap-up

In terms of the techniques being applied to do Sentiment Analysis, the research is divided into two camps, the supervised machine learning algorithms and the unsupervised lexicon based approaches. As for the supervised learning approach, the Naive Bayes algorithm is probably the most suitable algorithm to use for short texts like that of Twitter. The algorithms used should be trained and tested using test data from the same domain as the data to be analysed. The training and test data must be a representative sample from the data to be sampled. The algorithm features need to be tuned to find the optimum set of features that provide the best accuracy for the domain being analysed. One must be aware of bias in the training and test data, and take steps to correct for this, since this bias will affect the accuracy. An accuracy of 75% seems to be the baseline for Sentiment Analysis tools using machine learning algorithms at this point in time. As for unsupervised lexicon based approach, the literature review has shown that this technique can be as accurate as supervised approaches.

Sentiment Analysis, in itself, is not enough of an indicator for managers to be able to develop an effective public relations strategy. An analysis of the potential influence the opinion has (Influencers) must also be done. In order for a business to fully benefit from analysing social media streams they need to have an understanding of the “Influential Opinion” of their customers.

Managers should understand that not all of their customers are using social media channels and that they need to continue interacting with these
customers using the same techniques that they have been using. Establishing the correct balance is critical. This balance needs to be adjusted as the social media maturity level of their customer base changes.

What has also become evident is that the techniques being employed to do Sentiment Analysis are in their infancy, provide relatively low levels of accuracy and still need more research and development to maximise their value, however the potential of being able to have an insight into how the majority of a population feels, thinks, experiences, learns and interacts, is so alluring, that the concept of Sentiment Analysis cannot be dismissed. Being aware of the potential pitfalls of the techniques and then taking these into account when doing Sentiment Analysis can maximise the accuracy of the insights gained.

3. Influential Opinion analysis

The insights gained from the review were instrumental in the selection of the techniques used to do the analysis. For the analysis a lexicon based Sentiment Analysis tool was chosen. This approach is highly regarded and gives accurate results compared to other currently available techniques. The review revealed that measuring influence on follower count alone is not a credible method of measuring influence, this is accounted for in the influence analyses section. Finally the review showed that value can be derived from analysing social media streams, especially during significant events, as in the BP case study. This proved invaluable as in the analysis done, a similar, albeit less significant event, was identified for one of the companies selected.

The analysis presented in this section was done in order to firstly, find a sample of companies that have a high degree of interaction and secondly, to choose 3 companies from that sample and analyse the quality of these interactions and the potential those interactions have to influence others. In Section 3.2 the degree of interaction is measured by finding Tweets containing a sample of South African company names and a sample of popular South African brand names. In Section 3.3 an analysis of the Sentiment measured for 3 South African companies is provided. In Section 3.4 an analysis of the potential influence of a sample of Tweets for the 3 companies is provided.

The insights gained from the analysis helped to provide insight into the issues mentioned in the introduction of this paper, those being, understanding the degree of usage of social media for a business and for South Africa, monitoring competitors and comparing their results to one’s own and understanding who the influencers in one’s social media ecosystem are. The insights also helped to answer the question as to whether there is sufficient influential opinion being expressed to warrant South African companies continuously monitoring the social interactions of their customers in Twitter.

3.1. Acquiring the Twitter dataset

The Twitter search API was used to retrieve Tweets over the period 2012-06-15 to 2012-07-16. The Twitter Search API geocode filter parameter was set to -29.554345, 26.037598, 1000km. This seemed to be the best point and radius to use to ensure coverage of every part of South Africa. It did mean that Swaziland, Lesotho, the cities of Gaborone in Botswana and Maputo in Mozambique are also included in the area covered. The inclusion of these additional areas did not impact the overall results. The Twitter Search API was polled repeatedly, retrieving 100 Tweets at a time. The average Tweets collected per day were 715 305. The average excludes the 15/06/2012 and the 15/07/2012 as not all Tweets were collected on these days. The number of days Tweets were collected for is 32, with 30 of these being complete daily datasets. The total number of Tweets collected for the period was 21 646 136. Figure 1 shows a plot of the total Tweets collected each day. Twitter returns data in the JSON format. JSON, the JavaScript Object Notation
standard allows for data interchange between systems in a human-readable format. The ServiceStack library was used to parse the JSON files.

Although the frequency analysis was of little use it did lead to discovering a better way of tokenising Twitter texts. A new tokeniser was created using a regular expression tokeniser that proved effective in tokenising Tweets. Using this tokeniser a list of the top 10 hash-tags and user-mentions for each of the sentiment ratings -1 to -5 and +1 to +5 could be generated.

This is a possible alternative way, as opposed to using keywords, of identifying the relevant Tweets of a company or brand, but is very restricted as it only uses hash-tags or user-mentions. The error margin, in terms of identifying relevant Tweets is likely to be lower for this kind of analysis as the user-mention is unique to each Twitter user and hash-tags tend to be a unique word used by the Twitter community to identify specific topics. An example of using this kind of analyses is shown in

Table 1. The username @garethcliff was analysed for the number of times the token appears in the dataset as well as the sentiment sum for each sentiment range found.

<table>
<thead>
<tr>
<th>Sentiment Range</th>
<th>Number of times this token appears in Tweets</th>
<th>Sentiment Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>2746</td>
<td>-2746</td>
</tr>
<tr>
<td>-3</td>
<td>686</td>
<td>-2058</td>
</tr>
<tr>
<td>1</td>
<td>5514</td>
<td>5514</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>104</td>
</tr>
<tr>
<td>Total positive</td>
<td>5618</td>
<td></td>
</tr>
<tr>
<td>Total negative</td>
<td>-4804</td>
<td></td>
</tr>
</tbody>
</table>

Negative as a percentage of Total 46.09%

### 3.2. Identifying companies or brands with a high degree of interaction

#### 3.2.1. Word frequency analysis

Initially it was thought that a frequency analysis of the words in all the Tweets would be a viable way of identifying which businesses or brands were mentioned most often. While this was an interesting exercise the results demonstrated that it was in fact a naive approach as businesses or brands can be known by more than one name or by abbreviated forms of its name.

Although the frequency analysis was of little use it did lead to discovering a better way of tokenising Twitter texts. A new tokeniser was created using a regular expression tokeniser that proved effective in tokenising Tweets. Using this tokeniser a list of the top 10 hash-tags and user-mentions for each of the sentiment ratings -1 to -5 and +1 to +5 could be generated.

This is a possible alternative way, as opposed to using keywords, of identifying the relevant Tweets of a company or brand, but is very restricted as it only uses hash-tags or user-mentions. The error margin, in terms of identifying relevant Tweets is likely to be lower for this kind of analysis as the user-mention is unique to each Twitter user and hash-tags tend to be a unique word used by the Twitter community to identify specific topics. An example of using this kind of analyses is shown in

### 3.2.2. Business / brand keywords

Since the purpose of this exercise was to find comparable companies or brands with similar Tweet attributes like number of Tweets and distribution of positive and negative sentiment in the Tweets it was felt that using the frequency analysis technique mentioned above was too limiting as not all relevant Tweets would be found this way.

A list of publicly traded companies in South Africa was obtained from the JSE web site. There were 334 companies in the list. A list of brand names was obtained from the Sunday Times Top Brands

---

3 [http://Twitter.com/time](http://Twitter.com/time)

4 [http://www.jse.co.za/How-To-List/Main-Board/Main-Board-Listed-companies.aspx](http://www.jse.co.za/How-To-List/Main-Board/Main-Board-Listed-companies.aspx)
Survey\(^5\). This survey lists many of the most popular brands in South Africa. There were 126 brands in the list. Each item in the list was manually annotated with other known names for the company or brand. Some keywords appear in both lists as the brand name may also be the company name. The top 25 and bottom 25 ranked companies and brands are shown in Table 2 and Table 3.

<table>
<thead>
<tr>
<th>Brands</th>
<th>Companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>29879 time</td>
<td>7679 vodacom</td>
</tr>
<tr>
<td>16046 blackberry</td>
<td>5937 mtn</td>
</tr>
<tr>
<td>7949 lg</td>
<td>2576 absa</td>
</tr>
<tr>
<td>7679 vodacom</td>
<td>2422 woolworths</td>
</tr>
<tr>
<td>5937 mtn</td>
<td>1993 spur</td>
</tr>
<tr>
<td>5619 nokia</td>
<td>1999 aquarius</td>
</tr>
<tr>
<td>5181 samsung</td>
<td>1688 shoprite</td>
</tr>
<tr>
<td>5002 kfc</td>
<td>1665 spar</td>
</tr>
<tr>
<td>3895 nandos</td>
<td>1552 standard bank</td>
</tr>
<tr>
<td>3863 mcdonald's</td>
<td>1513 mr price</td>
</tr>
<tr>
<td>3428 fnb</td>
<td>1496 pick 'n pay</td>
</tr>
<tr>
<td>3262 bmw</td>
<td>1207 telkom</td>
</tr>
<tr>
<td>2946 budget</td>
<td>904 dig</td>
</tr>
<tr>
<td>2576 absa</td>
<td>859 picks</td>
</tr>
<tr>
<td>2538 nike</td>
<td>810 nedbank</td>
</tr>
<tr>
<td>2516 toyota</td>
<td>765 sabmiller</td>
</tr>
<tr>
<td>1993 spur</td>
<td>741 se</td>
</tr>
<tr>
<td>1754 eskom</td>
<td>686 bowler</td>
</tr>
<tr>
<td>1686 shoprite</td>
<td>660 bidvest</td>
</tr>
<tr>
<td>1665 mtn</td>
<td>548 sawd</td>
</tr>
<tr>
<td>1952 standard bank</td>
<td>460 delta</td>
</tr>
<tr>
<td>1529 castec</td>
<td>405 sanlam</td>
</tr>
<tr>
<td>1508 dhl</td>
<td>372 old mutual</td>
</tr>
<tr>
<td>1496 pick 'n pay</td>
<td>370 capitec</td>
</tr>
<tr>
<td>1401 wimpy</td>
<td>316 impala</td>
</tr>
</tbody>
</table>

Table 2: Top 25 companies and brands ranked by number of relevant Tweets found

The words “blackberry” and “time” appeared in a large number of Tweets. The reason “blackberry” appears so often is that some Blackberry devices append the text “sent from blackberry”. The word time can be used in many contexts besides mentioning the brand Time. In order to overcome this future research should allow for not only words to be found in the Tweets but also exclusion words, words / phrases that when found will exclude the Tweet from the results.

In contrast to this the companies ranked in the bottom 25 all have a single Tweet in the 30 day period. For companies like this, Twitter would not be a suitable social media channel to monitor for Influential Opinion unless they took active steps to promote the use of Twitter by their customers.

<table>
<thead>
<tr>
<th>Brands</th>
<th>Companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>71 kelvin</td>
<td>1 lufax</td>
</tr>
<tr>
<td>71 all gold</td>
<td>1 sephaku</td>
</tr>
<tr>
<td>65 brutal fruit</td>
<td>1 sa corporate real estate fund</td>
</tr>
<tr>
<td>59 europcar</td>
<td>1 kelly group</td>
</tr>
<tr>
<td>58 kimberly spin</td>
<td>1 roffes</td>
</tr>
<tr>
<td>56 ferrero rocher</td>
<td>1 securedata</td>
</tr>
<tr>
<td>53 protex</td>
<td>1 south ocean holdings</td>
</tr>
<tr>
<td>47 elastic</td>
<td>1 rubex</td>
</tr>
<tr>
<td>45 clover crush</td>
<td>1 lewis group</td>
</tr>
<tr>
<td>42 hollard</td>
<td>1 consolidated infrastructure</td>
</tr>
<tr>
<td>37 city lodge</td>
<td>1 central rand gold</td>
</tr>
<tr>
<td>35 postnet</td>
<td>1 orb holdings</td>
</tr>
<tr>
<td>33 financial mail</td>
<td>1 cape empowerment</td>
</tr>
<tr>
<td>30 logosun</td>
<td>1 bk one</td>
</tr>
<tr>
<td>28 equi-fort</td>
<td>1 conduct capital</td>
</tr>
<tr>
<td>22 hla-soft</td>
<td>1 hospitality property</td>
</tr>
<tr>
<td>21 savanna dry</td>
<td>1 fairest</td>
</tr>
<tr>
<td>14 salticax</td>
<td>1 delrand</td>
</tr>
<tr>
<td>13 white star</td>
<td>1 bell equipment</td>
</tr>
<tr>
<td>12 yardley</td>
<td>1 hulamin</td>
</tr>
<tr>
<td>10 mail &amp; guardian</td>
<td>1 country bird</td>
</tr>
<tr>
<td>9 pro vita</td>
<td>1 allied electronics</td>
</tr>
<tr>
<td>7 sasol</td>
<td>1 limited</td>
</tr>
<tr>
<td>6 spar</td>
<td>1 sasol</td>
</tr>
<tr>
<td>5 pick 'n pay</td>
<td>1 nandos</td>
</tr>
<tr>
<td>5 picknpay</td>
<td>1 shoprite</td>
</tr>
<tr>
<td>5 shoprite</td>
<td>1 checkers</td>
</tr>
<tr>
<td>5 checkers</td>
<td>1 shoprite checkers</td>
</tr>
<tr>
<td>4 spar</td>
<td>1 super spar</td>
</tr>
<tr>
<td>4 kwiskspar</td>
<td>1 spar</td>
</tr>
</tbody>
</table>

Table 3: Bottom 25 companies and brands ranked by number of relevant Tweets found

Pick n' Pay, Shoprite and Spar were selected as the companies or brands to further analyse. They are all in the same industry, and are therefore comparable. The search algorithm was modified to search for the keywords within the Tweets without requiring the keyword to appear as a complete word within the Tweet. Tweets where the keywords appear as substrings of the Tweet were accepted as relevant Tweets. This was done in order to find more relevant Tweets, however, this increased the risk of false positives. The keywords were modified to minimise this risk. The keywords used are shown in Table 4.

Table 4: Search keywords used

The number of Tweets found for each was similar. It was felt that the number of Tweets is of...
sufficient quantity. The group of companies have an average of 89 Tweets per day. The number of Tweets found for each of the selected companies is shown in Table 5.

<table>
<thead>
<tr>
<th>Business</th>
<th>Total Tweets</th>
<th>Average per Day (30 days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick n' Pay</td>
<td>3423</td>
<td>114</td>
</tr>
<tr>
<td>Shoprite</td>
<td>2746</td>
<td>92</td>
</tr>
<tr>
<td>Spar</td>
<td>1864</td>
<td>62</td>
</tr>
<tr>
<td>Average</td>
<td>2678</td>
<td>89</td>
</tr>
</tbody>
</table>

Table 5: Total Tweets for companies selected

3.3. Determining and analysing sentiment

Professor Michael Thelwall, the author of the SentiStrength library agreed to share the code for his library. It was refactored and used directly in the JSON processing code. Since we were mainly interested in opinionated Tweets the list of Tweets were filtered for each company. Tweets with sentiment = 0 were excluded. Since the first and last days of the dataset were not complete they were excluded as well. The numbers of Tweets after filtering is shown in Table 6:

<table>
<thead>
<tr>
<th>Business</th>
<th>Total Tweets</th>
<th>Average per Day (30 days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick n' Pay</td>
<td>1910</td>
<td>64</td>
</tr>
<tr>
<td>Shoprite</td>
<td>1266</td>
<td>42</td>
</tr>
<tr>
<td>Spar</td>
<td>1026</td>
<td>34</td>
</tr>
<tr>
<td>Average</td>
<td>1401</td>
<td>47</td>
</tr>
</tbody>
</table>

Table 6: Total Tweets for companies selected excluding sentiment = 0 and first and last days Tweets

3.3.1. Shoprite and Computicket

After the initial analysis of Shoprite it became clear that an event took place on the 19th that generated significant Tweets and negative sentiment. This was attributed to the fact that Shoprite made the decision, via Computicket, to sell Lady Gaga concert tickets and many people where Tweeting about this.

An analysis of Tweet distribution and sentiment for Shoprite and Computicket is shown in Figure 2 and Figure 3. An analysis of the Tweet distribution and sentiment levels for each company is shown in Figure 4 and Figure 5.

Another search was conducted for Shoprite and Computicket, this time separating the results into 3 categories, Tweets mentioning Shoprite only, Tweets mentioning Computicket only and Tweets mentioning both Shoprite and Computicket. This was done in order to see if the negative sentiment directed at Computicket affected the Shoprite brand.
A qualitative analysis of the Tweets showed that on the 19th many people were complaining about the poor performance of the Computicket website. Shoprite made the decision to open up counter sales of tickets on the 20th. On the 19th many people tried to buy tickets online and were frustrated by the poor performance of the system.

"When @Computicket put me in an online queue of 48000 people for my @LadyGaga tickets and I’m like... [Link] by the user "Dylanjack".

Table 7 shows the number of Tweets found on the 19th for each of the 3 categories.

<table>
<thead>
<tr>
<th>Business</th>
<th>Total Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shoprite Only</td>
<td>1113</td>
</tr>
<tr>
<td>Computicket Only</td>
<td>3967</td>
</tr>
<tr>
<td>Shoprite and Computicket</td>
<td>74</td>
</tr>
</tbody>
</table>

Table 7: Total Tweets found on the 19th

Another issue that was brought up was the fact that on the 19th the tickets were only available online.

"RT @mybroadband: Lady Gaga online exclusivity is discrimination: Shoprite ‐ 87% of South Africa’s population do not have access to th... [Link] by the user "MyBroadband".

Interestingly on the 19th the sentiment score for Tweets containing the keywords Shoprite and Computicket is positive. A qualitative analysis revealed that this was caused by Tweets announcing that on the 20th over the counter ticket sales would be opened. The sentiment swings positive on the 20th as many people were able to get their tickets and the mood seemed generally more positive.

"Finally got my @ladygaga tickets! Thanks Shoprite &amp; @Computicket #LadyGagaSA @channel24" by a user "JeanEsterhuizen".

On the 20th the sentiment swings positive; however the positive sentiment seems to be more about the Lady Gaga concert then Computicket. The negative sentiment seems to be directed related to Computicket and Shoprite.

There does not seem to be evidence that suggests Shoprite’s long term brand image was damaged by this event as the sentiment quickly swings positive after the event. It is possible that not enough people know that Computicket is owned by Shoprite or that the joy of finally getting tickets to the concert counteracted the initial negativity toward the brand.

![](figure4.png) Figure 4: Tweets distribution over the sample period for each company

![](figure5.png) Figure 5: Tweets sentiment sum per day over the sample period for each company

3.3.2. Shoprite

Table 8 shows the average sentiment for the 3 companies over the Tweet sample period.
<table>
<thead>
<tr>
<th>Business</th>
<th>Total Tweets</th>
<th>Sentiment Sum</th>
<th>Average Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick 'n Pay</td>
<td>1910</td>
<td>921</td>
<td>0.48</td>
</tr>
<tr>
<td>Shoprite</td>
<td>1266</td>
<td>467</td>
<td>0.37</td>
</tr>
<tr>
<td>Spar</td>
<td>1026</td>
<td>452</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 8: Average sentiment for the period

A qualitative analysis of the negative sentiment Tweets for Shoprite was done which ignored the influence of Computicket. There were some re-occurring themes. The most prominent of these were:

- Customer frustration with long queues
- Annoyance at Shoprite’s seemingly unauthentic advertising
- Customer frustration with bad service.

Of the three companies Shoprite has the lowest average sentiment.

3.3.3. Spar

The higher than normal peak on the 24th can be attributed to the Spar Ladies Walk event. The overall sentiment of this event was moderately positive with some negative sentiment Tweets but these referred to how tired or difficult the event was for the individuals concerned.

The only negative sentiment Tweet related to the event was to about roads being closed. Perhaps in future Spar should Tweet about the road closures to a prominent traffic monitoring Twitter account in order to avoid negativity toward its brand who are inconvenienced by it.

A Tweet on the 27th about the Spar group offering a R100 000 reward for information on the shooting of a Jukskei Park manger was measured by the system as negative, however this event seemed to seen in a positive light by people, who felt that Spar was taking a stand against crime.

Since Spar has the lowest number of Tweets of the 3 companies, it is more susceptible to swings in sentiment compared to the others. Even so it ranks as second, compared to the other 2 companies, in terms of average sentiment.

3.3.4. Pick n’ Pay

The negative peak on the 14th can be attributed to a particular Tweet being re-Tweeted many times.

“ATTENTION ALL RUNNERS. DUE TO FLOODING RISKS ON THE ROUTE THE PICK n' PAY CAPE TIMES KNYSNA FOREST MARATHON HAS BEEN CANCELLED.” by user “oysterfestival”.

This Tweet was sent out at 3:29am. At 03:41am the same Tweet was sent by the user “PicknPay”. This demonstrated good co-ordination between the Knysna Oyster Festival organisers and their sponsor Pick n’ Pay. These actions probably prevented the situation occurring of many runners coming to the event who did not know it was cancelled.

Of the 3 companies Pick n’ Pay has the highest number of Tweets and the highest sentiment score.

3.4. Determining influencers

While knowing what the sentiments being expressed by one’s customers are is useful one should also try and understand what the potential influence of these Tweets is. In order to do this part of the analysis a sample of users were selected and their last 200 Tweets were retrieved from Twitter. Included in the dataset were user attributes like followers count, statuses count (number of Tweets made since account creation), re-Tweet count for each Tweet as well as the number of Tweets found. Since the Twitter system only allows access to the users last 7 days’ worth of Tweets, the number of Tweets found for some users was less than 200. (200 Tweets would only be retrieved if the user made at least 200 Tweets in the last 7 days.) The positive and negative sentiment was calculated for each Tweet and then an average calculated for the positive and negative.
These were then summed to provide an overall average sentiment for the users Tweets. The re-Tweet count was also averaged for the user’s Tweets. Table 9 shows the Tweeter details for each of the Tweets discussed above. The PigSpotter user is an informational account on Twitter. Users follow this account in order to get traffic updates. The user’s sentiment for the last 200 Tweets is slightly negative. The re-Tweet count is also very high. While follower count has been shown to not always be a reliable indicator of user influence, this user’s followers are following because they have an interest in a similar topic and in this case follower count can be seen as an important indicator of user influence. The mybroadband user has similar attributes. Both PigSpotter and mybroadband are likely to be highly ranked influencers.

The user JeanEsterhuizen’s sentiment is very slightly negative, and has a sizable number of followers however in this case this is not a reliable indicator as this user’s may simply be following many users who in turn reciprocate by following her. Her re-Tweet count is also very low. She will not likely be an influential user. This same argument can be applied to the users Jvbtrafficguy and Indulgence_Cafe. While Jvbtrafficguy has a sizable following and is supposedly an informational account he has a very low re-Tweet count and is unlikely to be an influencer. oysterfestival has a high re-Tweet count and is informational, therefore likely to be a highly ranked influencer. An interesting user is PicknPay, who, as one would expect, has a positive user sentiment. This user has a very high follower count, is informational, and yet has a low re-Tweet count. This seems to indicate that while people follow PicknPay because they have an interest in that topic, they do not feel that the Tweets PicknPay send out are important or interesting enough to be re-Tweeted. In this case Pick n’ Pay who maintains this account may not realise that their accounts activity, while keeping their customers informed, does not do a good job of influencing others, via re-Tweets. This is critical if Pick n’ Pay wishes to attract new customers via the influence of its existing followers.

### 3.5 Analysis insights

What is clear from the analysis is that while a system for finding relevant Tweets and calculating sentiment on them is extremely useful, it is also necessary to do a manual qualitative analyse on the Tweets found in order to get a sense what is causing the sentiment being calculated.

This qualitative analysis provides insights, not only by looking at the Tweet but by looking at the comments on a Tweet and the number of Retweets. By following the Tweet “trail”, one can gain useful insights into the customers of a brand.
company, insights that can be used to improve a service, improve a product, improve goodwill or improve brand image.

Constantly monitoring customer sentiment is important because even though each day may not provide deep insight, there will be days where something goes wrong and having the ability to quickly respond to the problem and prevent the commentary going viral is invaluable. Monitoring trends in the data is also necessary as one can determine if initiatives to improve customer service, new marketing campaigns or new product launches are being effective.

South African companies which have a sufficient degree of customer interaction in social media streams can derive some value from continuously monitoring these streams. This value is especially maximised when a significant event occurs, an event related to a decision made by the company or an operational issue the company faces when implementing a new product or service. This was demonstrated by the Shoprite-Computicket analysis. This event most likely damaged the short term reputation of the Computicket brand. The Shoprite brand was most likely not highly damaged by this event as it seems not many people are aware of the fact that Computicket is owned by Shoprite. Like the BP example in the review, Computicket did little to re-establish its credibility during the event, tweets they sent tended to be informational only. The long term Computicket brand was not likely highly damaged by this event as the joy people experienced when they finally were able to get their tickets probably elevated the perception of the Computicket brand in their minds. This situation is unique to Computicket and other companies who experience operational issues when dealing with their customers are more likely to suffer long term reputation damage.

By being able to identify influencers a business can respond to a potential negative event much more effectively. By being able to change negative perceptions that influencers have toward the company or brand a public relations initiative can far more quickly counteract negative perceptions.

Monitoring the social media stream would allow managers to quickly identify problems and then respond to them. These kinds of events can occur at any time, without warning, and so monitoring the social media stream on an adhoc basis would not be as beneficial as monitoring it continuously.

4. Conclusion

The analysis of the South African Twitter stream has revealed some implications for South African businesses thinking about continuously monitoring Influential Opinion.

This kind of Twitter analysis will probably not be as beneficial for all types of South African companies. Depending on the kind of social media presence the company has chosen to have will influence the kind of results different companies can have. If the business has chosen to use Facebook as their primary social media channel then they are likely to have less activity on the Twitter social media channel. Companies who deal directly with the public (e.g. retail, like Pick n’ Pay) will tend to have more Twitter activity than those who do not (e.g. heavy industry like Bell Equipment). Companies who do not have a lot of Twitter activity should not dismiss this kind of analysis, as in certain situations, like the Lonmin tragedy\(^6\), monitoring the sentiment of the public, ones workers, trade unions and political parties would be key in planning ones public relations.

In order to establish if there is sufficient sentiment being expressed in Twitter to continuously monitor this social media stream, managers must first decide on a strategy for mining Twitter data. The analysis can be done for the company or for brands the company markets.

\(^6\) [http://mg.co.za/multimedia/2012-08-16-violence-erupts-at-lonmin](http://mg.co.za/multimedia/2012-08-16-violence-erupts-at-lonmin)
A finer grained search will reveal customer sentiment on more subtle aspects of a specific brand of the business. A less granular search will reveal more general sentiment the public have regarding the business. Once this strategy has been decided on an initial search can be done in order to get a sense of the amount of sentiment being expressed. A similar search should be done for the competitors of the business. Keyword should be carefully selected to minimise any false positives. Any alternative names the business may have, or if the business is known by a colloquial name, should be included in the list of keywords.

If both the business and its competitors have a low amount of interaction on Twitter, it is possible that there are no strong Influential Opinions regarding the business or that the type of business is not something the public is sufficiently aware of or interested in to Tweet about. If the competitors of the business have more Influential Opinion being expressed than the business being analysed, then the possibility exists that the business has not made a sufficient investment in promoting their Twitter account and is missing an opportunity to gather valuable Influential Opinion from their customers.

If both the competitors and the business are receiving sufficient Influential Opinion and the competitors receive high levels of negative Influential Opinion then there is an opportunity for the business to capture market share by interacting with these dissatisfied communities of Twitter users and demonstrating to them that the business can better fulfil their needs. If the reverse is true, and the competitors receive high levels of positive Influential Opinion and the business receives high levels of negative Influential Opinion then the business needs to quickly rectify the things they are doing wrong, if they don’t their competitors will capitalise on this opportunity to draw customers away from their business.

Businesses with high levels of Twitter engagement will find that manually searching the Twitter stream will be labour intensive, prone to errors and subject to interpretation by the person doing the analysis. It would be beneficial to the business to purchase social media data analytic services. These services need to use methods that have been proven in research. They must not only measure sentiment but provide a measure of influence, which is based on an attribute or attributes proven in research, not the number of followers attribute. They need to operate in real time (i.e. as Tweets are made the service acquires and analyses them). They need to provide services such as the ability for the user to set thresholds in the Influential Opinion and be notified when these thresholds are exceeded. By being real time and allowing for thresholds means public relations employees of the business can respond quickly, giving them a high probability of preventing negative Influential Opinion going viral. The services should provide an API that allows the business to integrate the data analysis results directly into their own information systems.

In conclusion, there is value in embedding social media in the Sales and Marketing areas of the business. In Sales it can be used to monitor customer perception of products or services. In Marketing it can be used to monitor the effectiveness of marketing campaigns. In PR activities it can be used to counteract negative perceptions developing toward the business or brand.

5. Future research

Future research could be done on analysing the social media stream of channels other than Twitter in order to compare them and see if there is a correlation in terms of Influential Opinion between them. An experimental Influential Opinion model is proposed below. This model attempts to quantitatively measure Influential Opinion. This model needs to be tested and its accuracy measured.
5.1. Influential Opinion Model

Based on the influence analysis done above, an idea started to evolve as to how the influence of a set of Tweets could be automatically calculated. The model proposed makes 2 assumptions. The first assumption is that if a Tweeter Tweets a Tweet with a sentiment that is radically different from the sentiment of the Tweets they normally Tweet then there is a higher probability that this Tweet will have the capacity to influence others. (Bigonha, Cardoso, Moro, Gonçalves, & Almeida, 2011) used a similar concept. The second assumption is that a Tweet with a high re-Tweet count has a higher probability of influencing others. This model is experimental and is offered as a vehicle for further research.

5.1.1. Variable 1: Sentiment deviation from normal

The first variable used in the Influential Opinion model is a variable that is used to replace the Tweets sentiment score with a value that accounts for the difference between the user’s recent sentiment and the sentiment of the Tweet being analysed. In calculating the potential influence of the Tweet, more weight is assigned to Tweets where the sentiment of the user’s last 100 Tweets is different from the sentiment of the found Tweet, the greater the difference the greater the weighting. The weights were calculated to ensure the adjusted sentiment score is never greater than 10 or less than -10. Table 10 shows the weights used to adjust the sentiment of the Tweet.

5.1.2. Variable 2: Re-Tweet score

The second variable is the number of re-Tweets the user has. Ranges were determined and are used for the re-Tweet count. Table 11 shows the ranges used for re-tweet score. The ranges allow for a max score of 10 for the variable. The max was calculated by determining what the most re-Tweeted Tweet was and using that value as the max value for a weight of 10. The most re-Tweeted Tweet as of 26 September 2012 as reported by mediabistro\(^7\) was a Tweet by @justinbieber which was re-Tweeted 200 000 times.

<table>
<thead>
<tr>
<th>Users Sentiment Last 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
</tr>
<tr>
<td>0.9</td>
</tr>
<tr>
<td>1.8</td>
</tr>
<tr>
<td>2.7</td>
</tr>
<tr>
<td>3.6</td>
</tr>
<tr>
<td>4.5</td>
</tr>
<tr>
<td>5.5</td>
</tr>
<tr>
<td>6.4</td>
</tr>
<tr>
<td>7.3</td>
</tr>
<tr>
<td>8.2</td>
</tr>
<tr>
<td>9.1</td>
</tr>
<tr>
<td>10.0</td>
</tr>
</tbody>
</table>

Table 10: Weights used in influence of user sentiment calculation

<table>
<thead>
<tr>
<th>Min</th>
<th>Max</th>
<th>Re-Tweet Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>109</td>
<td>4</td>
</tr>
<tr>
<td>110</td>
<td>381</td>
<td>5</td>
</tr>
<tr>
<td>382</td>
<td>133</td>
<td>6</td>
</tr>
<tr>
<td>1334</td>
<td>4665</td>
<td>7</td>
</tr>
<tr>
<td>4666</td>
<td>16327</td>
<td>8</td>
</tr>
<tr>
<td>16328</td>
<td>57143</td>
<td>9</td>
</tr>
<tr>
<td>30000</td>
<td>200000</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 11: Weights used for re-Tweet score

5.1.3. Calculating the Influential Opinion score

The variables described above are calculated for each Tweet found. They are then multiplied together. The average positive Influential Opinion score is the average of all positive Influential Opinion scores found in a set of Tweets. The average negative Influential Opinion score is the average of all negative Influential Opinion scores found in a set of Tweets. The overall Influential Opinion Score is calculated by calculating the difference between the positive Influential Opinion score and the negative Influential Opinion score. This difference can then be compared to previous calculations to determine trends. The

\(^7\) [http://www.mediabistro.com/allTwitter/Twitter-most-reTweets_b29141](http://www.mediabistro.com/allTwitter/Twitter-most-reTweets_b29141)
positive and negative Influential Opinion scores are a measure of the potential of a set of Tweets to either influence positively or negatively respectively. A max possible score of 100 and a min possible score of -100 for the overall average is possible. A business would want to maximise the positive score and minimise the negative scores.

6. References


\_r=1