**Text mining of social media feeds to perform sentiment analysis for technology release or advertising campaign**

Peter Carew\*, Ali Akbar Jilani\*\*

\*Lecturer in Computing,   
Waterford Institute of Technology, Waterford, Co. Waterford  
Ireland (Tel: 353-051-302628; e-mail: pcarew@wit.ie).  
  
\*\*Student, Department of Science and Computing,   
Waterford Institute of Technology, Waterford, Co. Waterford  
Ireland (Tel: 353-085-212-6940; e-mail: 20078735@mail.wit.ie).

Abstract: Social Media Sentiment Analysis.

*Keywords:* Data Management, Data Analytics, Business Intelligence, Process Management, Process Improvement.

**Introduction:**

After the popularization of online social networks (OSNs), Sentiment Analysis has become an extremely popular tool for its application in several analytical domains especially the web and social media. Social media includes a variety of specific tools or applications, such as blogs, Twitter, online videos, social networks, and other online and electronic tools. Twitter itself has contributed to reshaping the web from a mere static repository to a dynamic forum (microblogging service) where users can publish their thoughts and opinions along with other types of “user-generated content (UGC)” on any topic of interest. This content carries valuable information particularly for applications that require analysis of public opinion on a certain topic. This study focuses on one such application where different large organizations tap into this resource as they try to understand public opinion (sentiment analysis) about their products that are either launched or about to be launched. Below semblance (Figure 1: Source Google Trends) depicts the growing trend on the topic of “Sentiment Analysis”. Textual Domain offers a unique proposition for SA.



Figure 1: Searches on Google for the query: ‘Sentiment Analysis’. This ﬁgure shows the steady growth on the number of searches on the topic, according to Google Trends, mainly after the popularization of online social networks (OSNs).

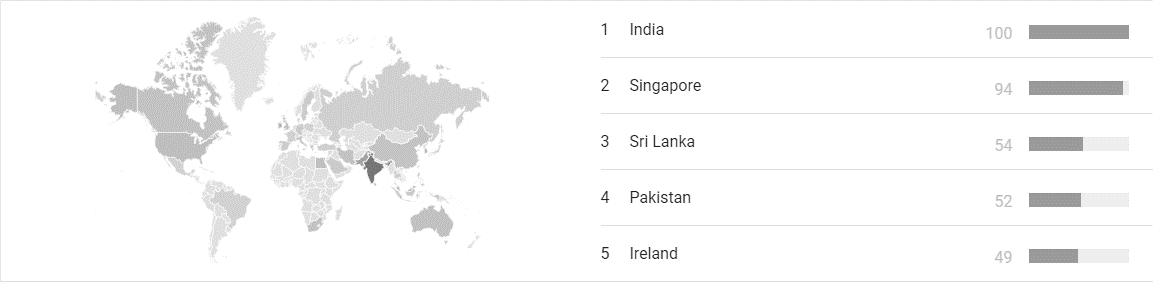


Figure 2: Top 5 ranking countries w.r.t. volume of search queries on 'Sentiment Analysis'.

According to Figure 2, most of the searches on Sentiment Analysis are found to be made from India while Singapore and Sri Lanka stand at second and third rank in terms of web searches on the subject. Below is the top 5 ranking of search volumes from different countries across the globe.

As Figure 3 shows, On the topic of ‘Sentiment Analysis’, ‘Twitter sentiment’ is the most common search term as per Google (Trends). The most common application of sentiment analysis is in the area of reviews of consumer products and services.

(Feldman, 2013) Defines ‘Sentiment Analysis’ or ‘Opinion Mining’ as the task of finding opinions of authors about specific entities. He explains how there is a huge explosion of ‘sentiments’ available from social media including Twitter, Facebook, message boards, blogs, and user forums. This opinionated information is a gold mine for companies and individuals that want to monitor their reputation or get timely feedback about their products and actions, may they be about product release. Sentiment analysis offers these organizations the ability to monitor the different social media sites in real time and act accordingly. Marketing managers, campaign managers, politicians, equity investors or even online shoppers can directly benefit from this sentiment analysis technology. (Feldman, 2013).

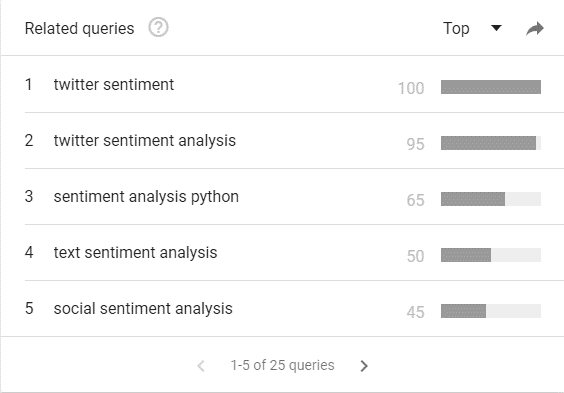


Figure 3: ‘Twitter sentiment’ is the most common query on the topic of 'Sentiment Analysis' as per Google Trends.

**Scope:**

*Why Twitter:* Sentiment Analysis is possible across the broad range of social media platforms available today. Below are some of the unique characteristics of twitter that distinguish it from other microblogging platforms such as Tumblr, FourSquare, Google+, and LinkedIn for sentiment analysis.

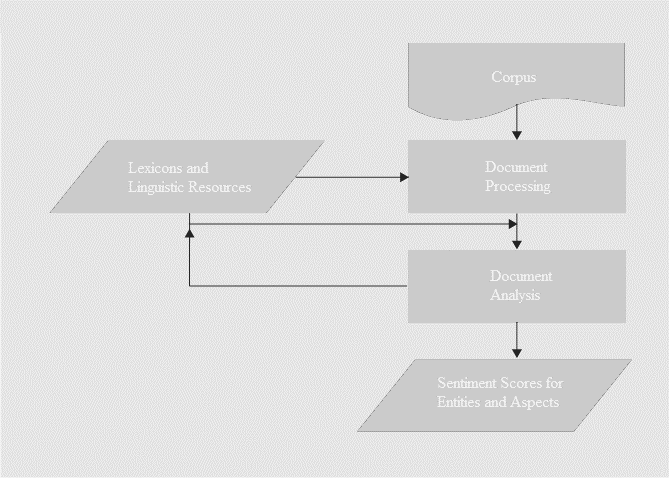
1. *Standard length:* Tweets have a standard length limitation of 140 characters which gives enough room to the Twitterati to explain his/ her opinion while remaining relevant to the topic.
2. *Informal type of medium:* Twitter seems to be the most suitable out of all other social media platforms as it offers an informal medium of expression (more suitable for subjective content) to its registered users while limiting them to 140 characters which helps control content relevance. Other microblogging platforms are either formal (LinkedIn) or are less popular than Twitter (Tumblr, Google+).
3. *Volume of content:* Over the years, Twitter's interface has remained simple, which is why a lot of tweets take place through third-party sites and applications that make the experience more useful. There could be other sources considered but volume and content relevance become important questions when you consider analyzing sentiments in products that are yet to be announced. In the context of technology release, there is a better chance of finding pre-release product centered content on twitter than any other social media platform also because of its popularity.

According to (Feldman, 2013), More than 7,000 articles have been written about sentiment analysis and various startups are developing tools and strategies to extract sentiments from text. The scope of this study is therefore being drifted more towards contribution to quality (as compared to variety) of results and to make it more manageable, it is also being confined only to *subjective sentences* (that contain opinions, beliefs and views) as opposed to *objective sentences* (that contain factual information). Subjective sentences carry the essence of sentimental information (opinions, beliefs and views) while objective sentences contain factual information that is more suitable to areas like stock picking. A Tweet may vary in the number of sentences it contains. These sentences may carry different opinions about the same entity. In order to develop an accurate and fine-grained view of different opinions, the proposed tool is required to attach sentiment annotations to individual sentences within a tweet. However, In order to limit the scope of research, following assumptions are being made.

* That tweets are written in English language. Re-tweets are excluded from the analysis.
* Since a tweet comprises up of more than a sentence, it may be assumed that the entire tweet contains an opinion on one main object expressed by the “Twitterati” (more reasonable in the context of document-level sentiment analysis)
* That we know the identity of the entity discussed in the sentence.
* It is assumed that each phrase in the sentence also contains just one opinion.
* To further relax the situation, it is assumed that there is a single opinion in each sentence.

**Research Approach:**

Figure 4: General architecture of a lexical Sentiment Analysis System.



There are multiple methods for measuring sentiments, including lexical-based and supervised machine learning methods. The study is designed to be conducted through development of a software tool that will gather relevant data through a supervised process. The supervised approach assumes that there is a finite set of classes into which data should be classified and training data is available for each class.

The scope for data gathering will be guided through filters applied to eliminate or isolate all hashtags that allow data more dispersed in content characteristics.

**Sentiment Analysis Tool:**

A generic sentiment analysis tool may employ a variety of linguistic techniques such as stemming, tokenization, part of speech tagging, entity extraction, and relation extraction used for pre-processing textual artefacts or the target subject (e.g. a Tweet). The tool may also utilize lexicons or other linguistic resources. The primary component in the tool is Document Analysis module that attach sentiment annotations to the subject (document, sentence or aspect of entity) by utilizing linguistic resources. These sentiment annotations being the primary output of the system may be presented to the user using a variety of visualization tools (Feldman, 2013).

(Peng , 2014) Invented a method for Social Media Sentiment Analysis. Accordingly, embodiments generally relate to systems and methods for generating a sentiment dictionary and calculating sentiment scores of adjectives within the sentiment dictionary. A set of seed words can be identified and expanded using synonyms and antonyms of the set of seed words. Social media data can be parse to identify adjectives that link to the set of seed words with the words "and" or "but." Matrices representing the attraction and repulsion among the linked adjectives can be generated. A factorization algorithm can be minimized to determine an output matrix that comprises positive and negative sentiment scores for each of the adjectives. In embodiments, a sentiment score for part of all of the social media data can be calculated using the output matrix, and one or more parts of the social media data can be classified as a positive or negative sentiment.

Sub Themes: Keywords for research papers search.

1. Social Media, Online, Web2.0
2. Sentiment Analysis in social media, Cyber-ethnography (Lee Sproul, Sarah Chrysler)
3. Data Mining and Text Mining
4. Technology Release, Digital Marketing

JISC (2010) defines social media or Web 2.0 technologies as “innovative online tools designed to enhance communication and collaboration”. Wikipedia, itself an example of social media, defines social media as “media for social interaction, using highly accessible and scalable communication techniques” that can also be thought of as “user-generated” or “consumer-generated content” (http://en.wikipedia.org/wiki/Social\_media).

Title,

Abstract,

Introduction,

Research Questions:

1. Is it possible to develop/ identify a tool or technology to predict success of technology release based upon sentiment analysis performed over pre-release historical social media data?
2. What would such a tool/ technology look like? (the tool should only proclaim authority if developed. If not, should not be a liability)

Conclusion or discussion of results.

Tables or Diagrams make it easy to understand..

**Ethical issues: was considered.**

Privacy tweet is not an issue in tweets.. open to words. People can use pseudo-names but they choose to use their identity instead. The study is designed around sentiment analysis of a particular subject and not an individual user. Group level content is what the tool am more interested in group level not individual level .. more aggregate level. General sentiment.. on hipe surrepounding technology release. negative sentiments. Little scope of misuse.

**Approach**:

Supervised and unsupervised sentiment analysis

Breakdown of sentences into Phrases.

Different types of sentences:

Conditional sentences, Question sentences, sarcastic sentences.

Before identifying the polarity (i.e., positive or negative) of a message, it is required that we analyze only subjective sentences and not objective

**GOAL!!!!**

**Proposal should answer these questions**

**1. What did you set out to do?**

**2. How did you set out to do it?**

**3. Why did you choose that approach (think about the RQs**

**and the theory)?**

**4. What methods were associated with your approach?**

**5. How did the approach fit the research questions?**

**6. Have I covered all the topics I need to cover in order to**

**explain my method?**

**Tool** – that is open source and well documented.

Literature Review: 12-15

A broad based literature review was conducted to find answers to below questions.

**- What Social Media sources have impact or are relevant can be used as sources of research data?**

**- What Social- Media/ Twitter Sentiment Analysis are commonly being used?**

**- What techniques/ algorithms are being used by people for sentiment analysis over twitter feeds?**

**Be ready for questions as to explain the algo how it works.**

**What results say..**

Methodology: (Any decisions on choices)

- I chose Algo X because … A.B.C.have used it.. backup your choices..

**- Why Twitter? (enables access through API to short messages/ Tweets that are 140 characters long and made up of plain text as compared to Facebook that is one page long and multi-media content type - not text based), Why not Facebook or WhatsApp?**

to test and compare lexicon-based and machine learning approaches to sentiment analysis, as well as their combination, using the LIWC2015 lexicon and RTextTools machine learning package. (Dhaoui C, Webster C, Tan L. Social media sentiment analysis: lexicon versus machine learning. *Journal Of Consumer Marketing* [serial online]. October 2017;34(6):480-488. Available from: Business Source Complete, Ipswich, MA. Accessed November 10, 2017)

**- How much of pre-release Twitter historical data will be required to perform analysis with?**

**- What Technology Release event would be suitable? (Windows 8, iPhone 8)**

1. Is it possible to develop a software tool or technology to predict success of technology release based upon sentiment analysis performed over pre-release historical social media data?
2. What would such a tool/ technology look like?

**1. Define the concepts in your research questions – refine the questions if**

**Necessary**

1. **Software Tools Development**
2. **Technology Release**
3. **Sentiment Analysis**
4. **Social Media**
5. **Software Architecture**

**2. Write down propositions which state relationships between the concepts**

**Propositions:**

R1. Success or failure of technology release are measureable through social media Sentiment Analysis.

R2. There is enough valuable pre-release historical data in social media that we can use for sentiment analysis

R3. Results of the S.A. Tool will be trust worthy.

R4. Architecture of the Software Tool will allow sentiment analysis of social media data to an acceptable level of accuracy. (S.A. Tool will be able to clearly analyze either success or failure of a technology release to a certain level of accuracy)

R5. Technology Release will be preceded by sentiment analysis

R6. It will be possible to perform S.A. through social media data

**Variables along top header and tests as rows.**

**Linux, Windows, Unix**

**what that data might look like..**

1. Cyber-ethnography (Lee Sproul, Sarah Chrysler)

Ethnography: living with eskimos to understand how they daily operate. How they stay warm, how they hunt. etc.

Participate onforums, own personality of .. emmerce yourself in to the life on internet.. tend to be done on a longer period of time..

Title:

Abstract:

-Background (Research Topic / Main problem being addressed)

10% -Research Questions (Very very clear) or if not clear enouph (Research Objective Statement.. optional)

-Scope of research Program (platform, technique)(Boundaries one approach is to use social media plateform use algos for S.A. for Irish Users.e.g. only negative sentiments.. sub-section of a problem area.

Knowledge Contribution: optional

~20% - Literature Review optional / State of the Art (what is new in technology):

Related Work: min. 5 pages lit rev. litra background 1-2 pages (how you got there). more demonstration that you are researching on it. Can use a ready-made algorithm.. Lit. Review is “What”

30% - Quality of work.

Analysis, evaluation of results, conclusion 20%

Communication (documentation layout and style, writing quality, presentation)

Theory/ Working Hypothesis: they associate to RQ. But use research questions to inform hypothesis. i.e. by testing these hypothesis, you get the answers to your research questions.

H1: If office walls are soft colored, they are more productive.

H2: Natural light leads to productivity.

H3: Offices with windows are more productive.

e.g.

Choose S.A. Algo

H1: Algo X is more suitable than Algo Y for S.A. regarding S.M. S.A.

Test both Algos and compare results to Accept or reject H1. If evidence is strong enough, accept it else reject it. Way of framing things.

H2:

Plan/ Schedule going forward: Future Plan:

* Milestones
  + When will I have tool developed(key dates)
  + Literature review to complete by
  + Data Gathering to complete by

Survey Offices/ people

**Preliminary Literature Review:**

A benchmark comparison (apple-to-apple) of twenty-four popular sentiment analysis methods (called the state-of-the-practice methods) across eighteen labeled datasets, originating from messages posted on social networks, movie and product reviews, as well as opinions and comments in news articles was conducted as they are used in practice, across multiple datasets (Ribeiro et al., 2016)

**Sentiment Analysis Algorithms:**

There are multiple methods for measuring sentiments, including Classification method (supervised machine learning) and lexical-based method.

*Classification Method Algorithms:*

(Feldman) Classification algorithms used Given the training data, the system learns a classification model by using one of the common classification algorithms such as Support Vector Machine (SVM), WSVM, C4.5 tree, AdaBoost, Linear kernel SVM, Naïve Bayes (NB), MaxEnt, Multi Naïve Bayes (MNB), CRF, Perceptron with Best Learning Rate, Voted Perceptron, Ensemble Method, Logistic Regression, or kNN.

*Lexicon Based Algorithms:*

SentiStrength, SentiCircles, Clustering-based word sense disambiguation (WSD), Lexicon-based classifier, Rule-based and ESSA.

This classification is then used to tag new documents into their various sentiment classes.

As per (Giachanou and Crestani) these characteristics themselves pose challenge to sentiment analyst.

(Bermingham and Smeaton) have concluded that classifying tweets is easier than classifying longer documents such as blogs.

**Challenges in sentiment analysis:**

(Giachanou and Crestani) have explained below characteristics of twitter as the main challenges faced by sentiment analysts.

1. *Text Length:* have explained tweet length limitation (140 characters) and informality of medium as challenges.
2. *Topic Relevance:* many researchers of twitter sentiment analysis have been considering presence of a word in a tweet as an evidence of topic relevance while others studies consider the hashtag symbol as a strong indicator or topic relevance. To a certain degree, these approaches may be correct as commonly the sentiment does target the topic.
3. *Incorrect English:*  Length Limitation and informality of communication make the language used in tweets is very different than the one used in other geners (web, blog, news etc.)
4. *Data Sparsity:* Owing the large volume of incorrect English and misspelled words, tweets contain an extensive amount of noise called “Data Sparsity” that negatively impacts sentiment analysis. Another reason for this noise is the use of non-standard textual artefacts such as emoticons and informal language. (Jeong et al., 2017) have also mentioned emoticons (‘^^’, ‘:-D’) and onomatopoeic words (‘haha’, ‘blah’) as a type of noice.
5. *Compositional Sentiments:*  (Feldman) has expressed the need for better modeling of Compositional Sentiments. At sentence level, this means more accuracy is required in overall sentence sentiment calculation from sentiment-bearing words, the sentiment shifters and the sentence structure.
6. *Anaphora and Auto-Entity Resolution:* Typically in an informal mode of communication, a product may be referred to by multiple names within a context. Anaphora resolution refers to aspect extraction e.g. “battery life” and “power usage” both mean the same thing(Feldman).

**Technology Release**

(Jeong et al.) see social media as an emerging source of customer voice since it assumed the form of a channel for exchanging and storing consumer-generated, large-scale, and unregulated voices about products. The authors have proposed a 4 step opportunity mining (identification of product opportunities) approach based upon topic modeling and sentiment analysis of large-scale customer generated social media data using open APIs. Below are the different steps discussed in the approach.

1. Use topic modeling to identify latent product topics used by product customers in social media
2. Quantify the importance of each product topic.
3. Use sentiment analysis to evaluate satisfaction level of each product using sentiment analysis.
4. Use the opportunity algorithm that uses product topic importance and satisfaction to determine opportunity value and improvement direction of each product topic from a customer centered view.

As a case study, opportunity mining of Samsung Galaxy Note 5 has been described as performed through the use of AIChemyAPI included in IBM’s Watson platform.

**Resources:**

A number of sentiment lexicon are available to be used for sentiment analysis. Given below are a few references.

1. General Inquirer Lexicon: http://www.wjh.harvard.edu/~inquirer/ spread-sheet\_guide.htm.
2. Emotion Lexicon: http://www.purl.org/net/emolex
3. Financial Sentiment Lexicons: http://nd.edu/~mcdonald/Word\_Lists.html.
4. MPQA Subjectivity Lexicon: http://www.cs.pitt.edu/mpqa/subj\_lexicon.html
5. SentiWordNet: http://sentiword-net.isti.cnr.it/
6. Sentiment Lexicon: http://www.cs.uic.edu/~liub/FBS/sentiment-analy-sis.html

Originality/value

– This paper’s main contribution is to demonstrat the competitive intelligence via the consumer opinion mining of social media data. Researchers, business analysts, and practitioners can adopt this method of social media analysis to achieve their objectives and to implement practical procedures for data collection, spam elimination, machine learning classification, sentiment analysis, feature categorization, and result visualization.

**References:**

BERMINGHAM, A. & SMEATON, A. F. 2010. Classifying sentiment in microblogs: is brevity an advantage? Ireland, Europe: Association for Computing Machinery.

FELDMAN, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM,* 56**,** 82-89.

GIACHANOU, A. & CRESTANI, F. 2016. Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Computing Surveys,* 49**,** 28-28:41.

JEONG, B., YOON, J. & LEE, J.-M. 2017. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*.

PENG , E. A. May 13, 2014. *Systems, methods and devices for generating an adjective sentiment dictionary for social media*

*sentiment analysis: [US Patent & Trademark Office, Patent Full Text and Image Database]*. United States, CA, Sunnyvale. patent application 13/082,963.

RIBEIRO, F. N., ARAÚJO, M., GONÇALVES, P., ANDRÉ GONÇALVES, M. & BENEVENUTO, F. 2016. SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science,* 5**,** 23.

TRENDS, G. *Google Trends - Sentiment Analysis* [Online]. Available: <https://trends.google.com/trends/explore?date=all&q=sentiment%20analysis> [Accessed 19/11/2017].