Social Media Posts Sentimental Analyzer

By

GROUP 19-5

DEPARTMENT OF NETWORKS

SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY

A Proposal submitted to the School of Computing and Informatics Technology

For the Study Leading to a Project Proposal in Partial Fulfillment of the

Requirements for the Award of the Degree of Bachelor of Science in Software Engineering of Makerere University.

Supervisor

Dr. Ntanda Moses

Department of Networks

School of Computing and Informatics Technology, Makerere University

mntanda@cis.mak.ac.ug/kmntanda@gmail.com, +256-70293-8609

|  |  |  |
| --- | --- | --- |
| **NO** | **NAMES** | **RE NO** |
| 1 | ATWINE NICKSON | 15/U/4058/EVE |
| 2 | MUGABO AMUZA | 15/U/7798/PS |
| 3 | TURYAHIIRWA HAPPINESS | 15/U/19330/EVE |
| 4 | MULINDWA HENRY | 15/U/8235/EVE |

OCTOBER 2018

# Introduction

As internet is growing bigger, its horizons are becoming wider. Social Media and Micro blogging platforms like Facebook and Twitter dominate in spreading news and trending topics across the globe at a rapid pace. A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These topics generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications.

Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents. It is intended to serve as an application to understand the attitudes, opinions and emotions expressed within an online mention. The intention is to gain an overview of the wider public opinion behind certain topics. Precisely, it is a paradigm of categorizing conversations into positive, negative or neutral labels. Sentiment analysis and opinion mining, due to its social and commercial value, has become a very hot topic of research these days. On other hand online, social media has become a most significant mode of communication on Web 2.0.

Hence sentiment analysis and user opinion mining on online social media has a great social and commercial importance. On social media, sentiment analysis for twitter has remained primary focus of researcher due to its simplicity while Facebook has been less addressed.[1] Hence in this research a framework is proposed to analyze Facebook posts and comments for opinions and sentiments of the public. Many people use social media sites for networking with other people and to stay up-to-date with news and current events. These sites (Twitter, Facebook) offer a platform to people to voice their opinions. For example, people quickly post their reviews online as soon as they watch a movie and then start a series of comments to discuss about the acting skills depicted in the movie. This kind of information forms a basis for people to evaluate, rate about the performance of not only any movie but also other products and to know about whether it will be a success or not. This type of vast information on these sites can be used for marketing and social studies. Therefore, sentiment analysis has wide applications including emotion mining, polarity, classification and influence analysis.

In the last three years, sentimental analysis has become a hot trend topic of scientific and market research in the field of Natural Language Processing and machine learning[2]

Extracting the public opinion from social media text provides a challenging and rich context to explorer computational models of natural language processing.

Data analytics, particularly in the context of big data, has become the Mantra for the new Data Scientists. By applying analytics to structured and unstructured data and mining, that is extracting items of information which are important for business planning and execution; enterprises are changing the way they plan and make business decisions[1].

# Problem Statement

Sentiment analysis of reactions to social media posts is still a challenge. This is as a result of language grammatical error, some comments are vague due to the use of slangs, sentiment and subjectivity are quite context-sensitive, and, at a coarser granularity, quite domain dependent.

Due to the above-mentioned challenges, we aim at analyzing textural data and build a sentiment classifier tool that is able to label positive, negative and neutral sentiments for that data.

# General Objective

To develop a classifier to accurately tag social media reactions as positive, negative and neutral.

# Specific objectives

* To gather and analyze requirements for building the Social media posts sentimental classifier.
* To design a classifier to classify and enable one to view classification of textual data into positive, negative or neutral sections.
* To implement a prototype of the classification system.
* To test and validate the developed Social media posts sentimental classifier.

# SCOPE

Our project aims at analysis of reactions to social media posts on Facebook. The study is going to be carried out on reactions to posted articles by daily monitor, new vision, observer and red paper on Facebook.

The project will cover comments only in English. Comments in any other language will not be covered in this project.

The research development and implementation of this project is in the duration that is stipulated for the final year project development.

# Significance

The product will be used as backbone for some application developers.

Our project will enable organizations and companies analyze massive feedback from people’s reactions on social media.

Political parties may be interested to know if people support their program or not and this project will help solve this problem.[3] This would reduce the costs spent on employing people to analyze these reactions manually.

Social organizations may ask people’s opinion on current debates and the problem of analyzing these reactions can be solved with this project.[3]

Universities can use our tool to analyze student feedback and comments garnered either from their own surveys, or from online sources such as social media. They can then use the results to identify and address any areas of student dissatisfaction, as well as identify and build on those areas where students are expressing positive sentiments.[4]

Local government departments can gauge public sentiment towards their department and the services they provide, and use the results to improve services such as parking and leisure facilities, local policing, and the condition of roads using our tool.[4]

# LITERATURE REVIEW

With the increase in the population of blogs and social networks, opinion mining and sentiment analysis became a field of interest for many researches.

Sentiment Analysis is becoming a promising topic with the strengthening of social media such as blogs, networking sites etc. where people exhibit their views on various topics.[5]

We cover a review of document findings, facts, evidence and ideas from other researchers who have worked on topics that are related to this project at hand. We conducted an extensive review of the literatures related to Social media posts sentimental analyzer systems.

There has been a large amount of prior research in sentiment analysis, especially in the domain of product reviews, movie reviews, and blogs. Researchers have also analyzed the brand impact of micro blogging.[6]

J. Read [7] shows that using emoticons as labels for positive and sentiment is effective for reducing dependencies in machine learning techniques. The authors applied SVM (Support Vector Machine algorithm) classifier and Naïve Bayes classifier for implementation and CRF learners to classify sentiments at the sentence level and then investigated several strategies to determine the overall sentiment of the document.

The study used an emoticon technique as one of the ways of overcoming the domain, time and topic problems so as to find a much larger and diverse amount of general text. For example, users of electronic methods of communication have developed visual cues that are associated with emotional states in an attempt to state the emotion that their (users) text represents.

The emoticons were used in such a way that the author effectively marked up the text with an emotional state. The marked-up text was used to train the sentiment classifier for example a smile indicates generally positive text and a frown indicate a generally negative text.

The emoticon trained classifier performs well (up to 70% accuracy) when predicting the sentiment of article extracts from emoticon datasets.

We can apply this study on Facebook data(reactions). This is because people react to posts using emoticons. These emoticons reflect their emotions which need to be analyzed.

Text classification using machine learning is a well-studied field [8]. Pang and Lee [9] researched the performance of various machine learning techniques (Naive Bayes, maximum entropy, and support vector machines) in the specific domain of movie reviews.

Intuitions seem to differ as to the difficulty of the sentiment detection problem. However, it seems that distinguishing the positive from the negative is relatively easy for humans, especially to text categorization.

Indicators of the positive and negative sentiments were used so as to classify the various sentiment texts into categorizations of positive and negative. Bo Pang and Lillian Lee [10] used machine learning methods to implement the classifier such as Naïve Bayes approach, maximum Entropy and Support vector machines. To enhance the implementation, other bag-of-features were used to implement the machine learning algorithms through pre-defining a set of features that can appear in a document and predicting the number of times a word appear in a document.

Bo Pang and Lillian Lee were able to achieve an accuracy of 82.9% using SVM with a unigram model. However, the fact of individuals coming up with the best set of keywords to be used as indicators was non-trivial and we suggest that it is worthwhile to explore corpus based techniques, rather than relying on prior intuitions, to select good indicator features and to perform sentiment classification in general.

We can use machine learning techniques mentioned in the above research to design a classifier that can categorize sentiment text into positive, negative and neutral segments.

There are present trends of cryptocurrency by means of Text Mining and comparative study of different algorithms such as; capturing of data, Data Processing, Sentiment Classification, Sentiment Polarity and Visualization by different Techniques. With help of application interfaces (Application Programming Interface), social media data of Facebook and twitter data using API’s and procedure was captured and used as datasets.

Sentimental Analysis procedure for Present trends of Cryptocurrency and Comparative study of different algorithms such as Naïve Bayes, SVM was implemented for the classifier. The implementation also includes Classification methods for Sentimental analysis such as Machine Learning Based method, Lexicon Based method, Hybrid Based method and the features of sentiment classification methods with research background of machine learning approach, N-gram Sentimental Analysis, Ruled and Lexical approaches, Maximum Entropy Classifier, support vector machine and Naive Bayes Method. The conclusion was that the machine learning methods such as Support vector machine, Maximum Entropy and Naive Bayes Methods were more efficient and semantic analysis of text made process easier while doing sentimental analysis of social media.[10]

In [11], the authors use web-blogs to construct a corpora for sentiment analysis and use emotion icons assigned to blog posts as indicators of users’ mood.

The authors applied SVM and CRF learners to classify sentiments at the sentence level and then investigated several strategies to determine the overall sentiment of the document.

As the result, the winning strategy is defined by considering the sentiment of the last sentence of the document as the sentiment.

We cannot apply their method because they only considered last sentences. This leaves out the beginning and the middle sentences which may affect the classification of the whole data at large hence not giving the real attitude towards that post.

This [4] focused on performing effective Sentimental analysis and Opinion mining of Web reviews using various rule-based machine learning algorithms. SentiWordNet that generates score count words into one of the seven categories like strong-positive, positive, weak-positive, neutral, weak-negative, negative and strong-negative words was used. The proposed approach was experimented on online books and political reviews and demonstrates the efficacy through Kappa measures, which has a higher accuracy of 97.4 % and lower error rate. Weighted average of different accuracy measures like Precision, Recall, and TP-Rate depicts higher efficiency rate and lower FP-Rate. Comparative experiments on various rule -based machine learning algorithms have been performed through a Ten-Fold cross validation training model for sentiment classification. [4]. We can use this method to generate a general summary of number of people who react to the posts.

Micro blogging has become a popular communication tool among internet users.[12] Therefore, micro blogging websites are a rich source for opinion mining and sentimental analysis. Because micro blogging has appeared recently, there are a few research works that were devoted to this topic.

Pak, Alexander Paroubek and Patrick [12] focused on using twitter the most common micro blogging platform, for the task of sentimental analysis. They show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. They perform linguistic analysis of the collected corpus and explain discovered phenomena. Using the corpus, they build a sentiment classifier that is able to determine positive, negative and neutral sentiments for a document. Experimental evaluations show that their proposed techniques are efficient and perform better than previously proposed methods. In this research, they worked with English; however, the proposed technique can be used with any other language. Using Twitter API, they collected a corpus of text posts and formed a dataset of three classes: positive sentiments, negative sentiments, and a set of objective texts (no sentiments). The two types of collected corpora were used to train a classifier to recognize positive and negative sentiments. In order to collect a corpus of objective posts, they retrieved text messages from Twitter accounts of popular newspapers and magazines, such as “New York Times”, “Washington Posts” etc. They queried accounts of 44 newspapers to collect a training set of objective texts.

We can use the above study to design a classifier that is able to determine positive, negative and neutral segments for reactions to Facebook posts.

The impact of a politician’s speech can now be monitored real-time by the help of current studies[13]. For example, the researches on Arabic Spring and the effect of social media on the Tunisian case [13] or French Presidential Election and social media research [10] or Iran Green Movement from the twitter data [14] or research on UK 2010 election and effect of social media [15] are only a few researches on the topic.

In most of the researches, the data is collected from the social media like Twitter [9-12] or Facebook [16] or e-learning environments mixed with social networks [14]. All of these studies have a text mining part. Zhai [9] shows that the studies based on TF-IDF has a higher success than suffix tress or n-gram based approaches for Chinese case with the SVM classifier.

Some of the researchers prefer using the metrics built on the social network itself. For example, in Twitter, it is possible to get the number of followers and following and such information may be useful to calculate the political views of people depending on who they follow as in UK

Election research [15] where the feature extraction is built on the followers or following. Or on some other researches, text mining approaches like bag of words, interjection of emotics, part of speech tagging methods are implemented together.[17]

The TripAdvisor website is a lexical based system used to test if there are biases in lexical sentiment analysis accuracy between reviews authored by males and females. It is concerned with reviewing datasets of hotels and restaurants to contrast the accuracy of lexical sentiment analysis for males and females. However, the findings show that male sentiment is harder to detect because it is less explicit. This came about since there was no evidence that this problem could be solved by gender-specific lexical sentiment analysis.

Sentiment analysis is used as a black box solution by marketers who see the results of the algorithm used to classify sentiment but are not interested in its details. For example, they may find that 45% of comments about product A are positive in comparison to 25% for product B, concluding that product A is more favorably viewed. Some emotion-related terms, such as love and haha, are disproportionately used by one gender. The same is true for some other words and

linguistic features, such as exclamation marks.

The algorithm used here is **SentiStrength** which is a widely used lexical algorithm and

incorporates a machine learning element that can be used to adjust it to specific

tasks, such as for gender-specific sentiment analysis (see below).

Since the Twitter corpora were unrestricted by topic, it is possible that some words were used by males for one topic but by females for another (e.g.,

dogfighting) so that the improved performance could be due to differing topics rather

than methods of expressing sentiment.[18]

# Conclusion

The purpose of this review was to view past research made about text mining, sentiment analysis and available systems that enable users to do the sentiment analysis and to identify the gaps in the existing systems, their strengths and weaknesses. From this information found about other existing systems, we analyzed and got a clear understanding of what our proposed system will be comprised of basing on the weaknesses of those existing systems.

# References

[1] “a92367f4609c36a1f48751a7bc8fc8faac0e05fa @ conferences.unicom.co.uk,” 2014. [Online]. Available: http://conferences.unicom.co.uk/data-and-sentiment-analysis/.

[2] “5-things-sentiment-analysis-classification @ www.kdnuggets.com.” [Online]. Available: https://www.kdnuggets.com/2018/03/5-things-sentiment-analysis-classification.html.

[3] D. Virmani, V. Malhotra, and R. Tyagi, “Sentiment Analysis Using Collaborated Opinion Mining,” no. March, pp. 3–6, 1992.

[4] B. Angela, “b4f30ae34a9f06c0a763307cfa8d67b9a4d68978 @ typely.com.” [Online]. Available: https://typely.com/blog/making-use-of-sentiment-analysis/.

[5] O. P. Yadav and S. Ray, *Computational Intelligence in Data Mining*, vol. 711, no. December. 2019.

[6] B. J. Jansen, K. Sobel, M. Zhang, and A. Chowdury, “The Commercial Impact of Social Mediating Technologies : Micro-blogging as Online Word-of-Mouth Branding,” *Proc. 27th Int. Conf. Ext. Abstr. Hum. factors Comput. Syst.*, no. December, pp. 3859–3864, 2009.

[7] J. Read, “Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification,” no. June, pp. 43–48, 2005.

[8] C. D. Manning, “Foundations of Statistical Natural Language Processing.”

[9] B. Pang, L. Lee, H. Rd, and S. Jose, “Thumbs up ? Sentiment Classification using Machine Learning Techniques,” no. July, pp. 79–86, 2002.

[10] M. G. Bhargava and D. R. Rao, “Sentimental analysis on social media data using R programming,” vol. 7, pp. 80–84, 2018.

[11] H. S. Christopher D. Manning, “cs511-1.pdf.”

[12] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining,” p. 23.

[13] A. Younus, M. Azam, and M. Saeed, “What do the Average Twitterers Say : a Twitter Model for Public Opinion Analysis in the Face of Major Political Events,” pp. 618–623, 2011.

[14] K. K. Khonsari and A. Fathalian, “Social Network Analysis of Iran ’ s Green Movement Opposition Groups using Twitter,” pp. 414–415, 2010.

[15] S. E. Seker, “Sentimental Analysis on Turkish Blogs via Ensemble Classifier.”

[16] S. E. Seker, “Sentimental versus Impact of Blogs.”

[17] J. M. Martín, A. Ortigosa, and R. M. Carro, “SentBuk : Sentiment analysis for e-learning environments.”

[18] M. Thelwall, “Gender bias in machine learning for sentiment analysis,” *Online Inf. Rev.*, vol. 42, no. 3, pp. 343–354, 2018.