ISSN:2249-5789
Kavita Mainalli et al , International Journal of Computer Science & Communication Networks,Vol 5(2),96-102

Technique for Sentiment Analysis to Identify Social Disturbing Tweets Using Recursive Neural Tensor Network

Kavita Mainalli
BEC, Bagalkot
kavitam33@gmail.com

Prof. Savita.S.Hanji
BEC, Bagalkot
savitawali@gmail.com

Prof.M.M.Kodabagi
BEC, Bagalkot
mmkodabagi@gmail.com

Abstract

Sentiment analysis is a well known technique for finding sentiments from text data. Sentiment analysis on twitter grabbed the attention of researchers, as sentiment expressed by one user affects the sentiment of people involved in discussion via tweets. This is considered a harder problem due to the unique characteristics possessed by it. Thoughts expressed in such social media may affect reputation of product or movies or celebrities or social peace. Sentiment analysis is also one of the applications of Big data and can be used to determine polarity of these thoughts that is; whether it affects peace of society or not. This work is to classify large text data of twitter as it affects the society or not.

1. Introduction

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. In this research work tweets are the source material. It is used to determine attitude of tweeter user towards society. There are three levels of sentiment analysis. First, document level sentiment analysis, here basic information unit is a single document of opinionated text and single review about a single topic is considered. But in the case of forums or blogs, comparative sentences appear. Hence second level, sentence level sentiment analysis is performed. In sentence level sentiment analysis, the polarity of each sentence is calculated. The methods used for document level sentiment analysis can also be applied for sentence level sentiment analysis. In this level, objective and subjective sentences are identified. The subjective sentences contain opinion words which help in determining the sentiment about the entity. After which the polarity classification is done into positive and negative classes. Third level, phrase level sentiment analysis is a much more pinpointed approach to opinion mining. The phrases that contain opinion words are found and a phrase level classification is done.

The sentiment classification is also divided into two categories: binary sentiment classification and multi-class sentiment classification. Binary sentiment classification involves classifying sentiments either positive or negative. Multi-class sentiment classification involves classifying sentiments into one of five categories: strong positive, positive, neutral, negative and strong negative.

In earlier days most sentiment analysis work has been done on review sites [2][7]. Review sites provide with the sentiments of products or movies, thus, restricting the domain of application to solely business. The rise of social media like Twitter, Facebook, LinkdIn fuelled interest in using sentiment analysis to identify public opinions and interests. Among these social medias the twitter gained popularity because of its attracting features like platform for breaking news, not only meant for friends like facebook, any user can follow any other users of his interest, it is a great tool for brands to promote themselves and their products, high percentage of famous personality.

Several open source software tools utilize machine learning, statistics, and natural language processing techniques to automate sentiment analysis on a large collection of texts from twitter. The most common machine learning techniques used for sentiment classification include Naive Bayes, Maximum Entropy, and Support Vector Machine [8]. Most sentiment analysis algorithms use simple terms to express sentiment. However the cultural factors, linguistic nuances, and differing contexts prevent researchers from drawing the sentiment accurately.

Two main approaches to perform the sentiment analysis are bag of words approach and Deep Convolution learning. Compare to Bag of words approach Deep convolution approach using Recursive Neural Tensor Network out performs by giving 80.7% of accuracy [20]. Fig 1 shows example of RNTN approach. It also captures negation and its scope in sentence.

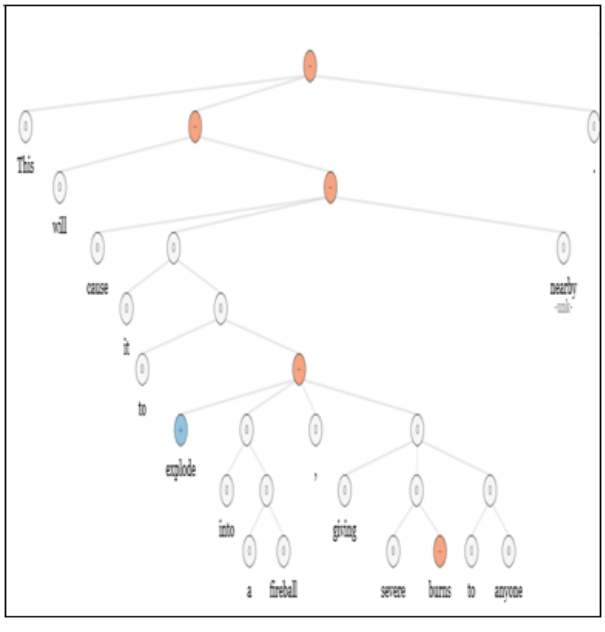

Fig 1: Example of RNTN

The input to RNTN is sentiment treebank built by Standford University on 11,000 movie related sentences. This study mentioned that extracting sentiment from the twitter short text is challenging task and that needs to be addressed. The proposed research work uses RNTN and Sentiment Treebank to understand the language representation of tweets, and also to classify tweets into five major classes (––,–, 0, +, ++). Short text are also handled along with identifying existence of social disturbing words in tweet before getting the sentiment of them.

2. Related work

Sentiment analysis of tweets is considered as a much harder problem than that of conventional text such as review documents. This is partly due to the short length of tweets, the frequent use of informal and irregular words, and the rapid evolution of language in Twitter. A large amount of work has been conducted in Twitter sentiment analysis following the feature-based approaches:

Ainur Yessenalina et al. [2] investigated structured models for document level sentiment classification. Presented 3 variation of SVM with respect to structure model they considered Sentiment classification with Latent Explanations model (Svm-sle),SVM prior model or standard model (SVMsle w/ Prior) and Feature smoothing model (SVM-sle-fs). Evaluations is performed on movie reviews and U.S. Congressional floor debates. SVM that is based on models proposed and trained on the Movie review data set results accuracy from the accuracy 92.29 to 92.50 and when it is trained on the U.S. Congressional Floor Debates dataset for 77.09 to 77.67.

As the usage of internet is grown, the people started using most of social networks like blogs, forums where people share their views on particular domain. Wojciech Gryc et al. [3] employ a hybrid machine learning and logic-based framework which operates along three distinct levels of analysis. It analyze political blogs data and train system using standard Naive Bayes Multinomial (NBM) and Logistic Regression models available in the WEKA11 toolkit.

Researches inspired by looking of the sites that allow people express their sentiment and opinion on multiple domains like twitter. Chenhao Tan et al [4] proposed Heterogeneous Graph Model with Direct estimation from simple statistics (HGM-NoLearning) to retrieve parameter, Heterogeneous Graph Model with SampleRank (HGM- Learning) to get user level sentiment labels, to identify opinions from social relationships using Twitter data set.

Oscar T¨ackstr¨om et al. [5] proposed semi-supervised latent variable models for sentence-level sentiment analysis and demonstrate how to combine coarse-grained and fine-grained supervision to benefit from sentence-level sentiment analysis – an important task in the field of opinion classification and retrieval.

Above work didn't consider multiple aspects of social media, to address this Erik Tromp [6] investigate automated sentiment analysis on multilingual data from social media. It presents a four-step approach to perform sentiment analysis. Approach comprises language identification, part-of-speech tagging, subjectivity detection and polarity detection. For language identification an algorithm LIGA which captures grammar of languages in addition to occurrences of characteristics. For part-of-speech tagging use an existing solution called the TreeTagger,

developed at the University of Stuttgart. Apply AdaBoost using decision stumps to solve subjectivity detection. For polarity detection proposed an algorithm RBEM which uses heuristic rules to create an emissive model on patterns.

Along with aspects of data requires one to decide upon which level of sentiment analysis will work for social data , Aurangzeb khan[7], proposed the rule based domain independent sentence level sentiment analysis method is proposed. The proposed method classifies subjective and objective sentences from reviews and blog comments. The semantic score of subjective sentences is extracted from SentiWordNet to calculate their polarity as positive, negative or neutral based on the contextual sentence structure. The results show the effectiveness of the proposed method and it outperforms the machine learning methods.

Bing Liu[8] presents discussion on all aspects of sentiment analysis and opinion mining like problems in SA, different level of SA and available different classifiers. Hassan Saif [9] different challenges that tweets poses and pilot work done on alleviating data Sparcity for twitter and also on political tweets sentiment analysis using NB classifier.

Discussions always based on particular topic, it may be based on the mood of people involved. Existing general purpose social web sentiment analysis algorithms may not be optimal for texts focused around specific topics Mike Thelwall et al.[10] introduces two new methods, mood setting and lexicon extension, to improve the accuracy of topic-specific lexical sentiment strength detection for the social web.

Anurag Mulkalwar et al. [11] proposed new approach to perform sentiment analysis on movie review called as Combined Approach. It uses two separate classifier Support Vector Machine (SVM) and Hidden Markov Model (HMM) as none of individual supervised or unsupervised classifier are not sufficient in achieving much precision. Then it combines results of these classifier using classifier combine rule. Data considered is movie review comments.

Diana Maynard et.al[12] describes the approach to the analysis of social media, combining opinion mining from text and multimedia (images, videos, etc), and cantered on entity and event recognition. It provides two main innovations: first, the novel combination of text and multimedia opinion mining tools; and second, the adaptation of NLP tools for opinion mining specific to the problems of social media.

Anders Westling [13] provides the approaches to understand how the user sentiment is expressed on twitter during event of crises and what kind of help they are expecting using machine learning approach.

To overcome the manufacturing feature engines as it is time consuming and not enough to capture the complex linguistic phenomena on microblogs Duyu Tang et al. [14] utilized pseudo labelled data, which is extensively explored for distant supervision learning and training language model in Twitter sentiment analysis, to learn the sentence representation through Deep Belief Network algorithm.

Xia Hu et al. [15] investigated whether social relations can help sentiment analysis by proposing a Sociological Approach to handling Noisy and short Texts (SANT) for sentiment classification. In particular, it present a mathematical optimization formulation that incorporates the sentiment consistency and emotional contagion theories into the supervised learning process; and utilize sparse learning to tackle noisy texts in microblogging.

M. Sakthivel et al. [16] presentd a study of different approaches to the state of the art techniques and current research in Sentiment Analysis based approaches for understanding user's context. Information about social relationships can be used to improve user-level sentiment analysis. It uses SVM classifier in SVM light package on data obtained by IMDb archive of the rec, arts, movies, reviews, news, and group. It contains 27,886 unprocessed and unlabeled HTML files that convey opinions of different authors on different movies.

The most recent work on sentiment analysis is presented by Anurag Mulkalwar et al. [20] explored new approach towards classification of sentiments which are present in textual content. To support Hidden Markov Model it suggests some transition rules for model rather than transition probability. It effectively uses part of speech tagging for formation of transition rules.

It is observed from literature survey that existing methods of sentiment analysis did not handle existence of short text in tweets. Also methods focused mainly on movie and product reviews data.

3. Motivation

Few real time scenarios motivated to initiate this work , such as, on April 23 2013, the official twitter account for the Associated Press, was hacked for a few hours, during which time a tweet was sent that said there had been a terrorist attack on the white house, and that President Obama had been injured. Luckily, the information was completely false, although the results of which produced a drop in the New York stock exchange of 138 points on the Dow Jones index (It is a benchmark for the health of the stock market).Another scenarios is Mehadi case, he created Twitter Account in different name and used to post prone tweets influencing negative thoughts in followers mind. This current work help society by identifying such activities quickly before it gets spread as virus all around the globe.

The existing methods of sentiment analysis did not handle existence of short text in tweets. Also methods focused mainly on movie and product reviews data. A new algorithm that identifies short text and replaces it with its meaning is developed. The proposed work is focused on obtaining sentiment of tweets that shows positive or negative attitude of user towards society.

4. Proposed model

Fig 2 shows the proposed model, which contains various phases starting from data acquisitions followed by pre-processing, representation, classification and action phase.

4.1 Tweet Collection

Collecting tweets from Twitter is very complicated job as twitter imposes many constraints on its developer. To collect tweets one has to create Twitter App under developer section as first step. Once app is created, will get developer Consumer Key and its secret key with which can contact twitter to obtain its services using Twitter4J API. The proposed solution is targeting all kind of users including the reputed personalities, so to obtain tweets from such personalities , here taken 32 college students having Twitter accounts to register to MyApp application by following persons of their own interest. Once the user authenticate MyApp will receive their respective accessToken and its seceret key with which can collect users timeline tweets and home timeline tweets for every 15 minutes via Quartz scheduler. So far 27,386 tweets are processed through model.

4.2 Pre-Processing

Using regular expression tweets are cleaned from URL and special symbols except below ones. While processing the tweets few of the special character are considered as they add meaning to sentence (**Table 1**).

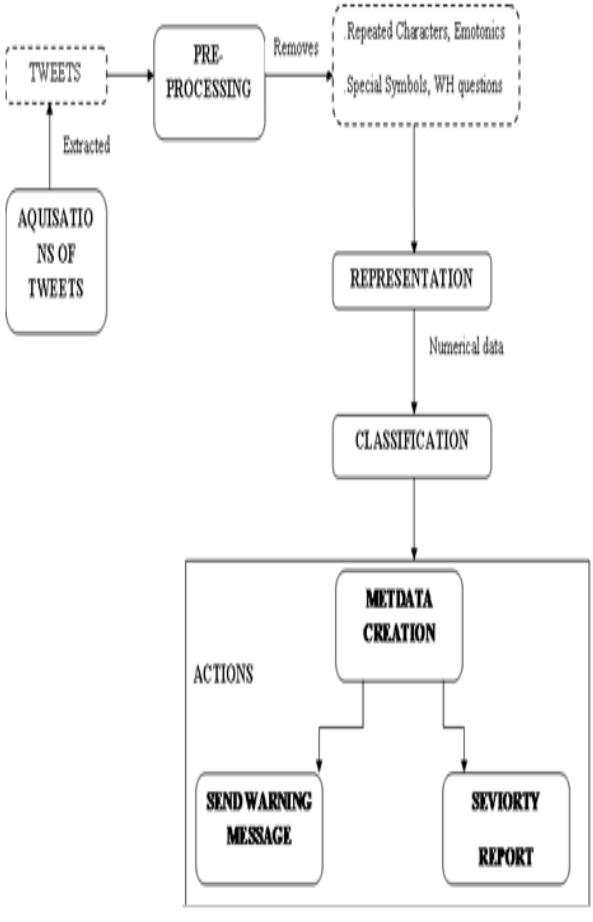

Fig 2: Proposed Model

Table 1: Retained Special Character

Symbol	Usage
@	Person, Place or Time
$	Price
#	Hashtag
_	Screen_Name
.	7.30
&	Connector

Cleaned tweets are stored as they required for the further analysis.

4.3 Short-Text Replacement

The main challenging task in tweet analysis is presence of irregular short text. The social disturbing content may present in such formats, so not to neglect short words , this work taken around 2500 globalized short text from www. netLingo.com . Processed tweets are checked for the presence of short text if any then it will be replaced by its respective meanings. These replaced tweets will be verified for existence of social disturbing words like kill, terrorist, abduction etc if any then such tweets are sent to Stanford sentiment analyzer.

4.4 Stanford Sentiment Analyzer

The Stanford analyzer is built on Recursive Neural Tensor Network RNTN proposed here [20], takes the input phrases and represent through word vectors and a parse tree, then compute vectors for higher nodes in the tree using the same tensor based composition function. Unlike bag of words techniques, it also accurately captures the sentiment by giving negative results for the negation of positive phrases. This study was focused primarily on semantic vector spaces, compositionality in vector spaces, logical form, deep learning and sentiment analysis. The sentiment of each tweets obtained from this analyzer will be stored into database.

4.5 Action Phase

In this phase severity is calculated based on difference between total tweets and total negative tweets. User activity and severity indicates users attitude towards society. If users activity is negative or strong negative and severity is high then warning will be posted on user timeline to stop such activities. This work also captures users personal details like email-id, user name , number of friends and location, these information may help cyber crime police to address such users.

4.6 Comparative Analysis

Comaparative analysis is performed between SEMANTRIA and StandFordNLP , among these standford gives more expected results compare to Semantria.Part of test cases are given in **Table 2.**

Table 2: Comparative results

TWEETS	EXPECTED RESULT	ACTUAL RESULT	STANFORD	SEMANTRIA
The government's main aim is to beat inflation	Neutral	Neutral	Passed	Negative
A variety of states and groups continue to seek to acquire weapons of mass **destruction** and the means to deliver them.	Negative	Negative	Passed	Passed
This will cause it to **explode** into a fireball, giving severe burns to anyone nearby.	Negative	Negative	Passed	Negative

5. Results and analysis

Collected tweets are maintained with mappings between tweet id , user id, date of creation and content of tweet. Part of this information is shown in **Table 3**. These processed tweets will undergo short text replacement , identification of social disturbing words and sentiment analyzer phase. Polarity of each tweets will be maintained with respects to its id as shown in **Table 4.** Based on this information severity and overall activity of user will be calculated. These results are of home timeline tweets means tweets posted by people whom the registered user is following. Fig 4 shows the activity report of registered user. These processed tweets will undergo short text replacement , identification of social disturbing words and sentiment analyzer phase. Polarity of each tweets will be maintained with respects to its id as shown in **Table 4.**

Table 3: Part of timeline_details content

Tweet Id	Original Tweet	Processed Tweet
569139627954741000	Ukraine's president says Putin aide was behind sniper killings of activists in Kiev protests http://t.co/GPnm2KJsYS http://t.co/FIruhCSAuz	Ukraine's president says Putin aide was behind sniper killings of activists in Kiev protests

Based on this information severity and overall activity of user will be calculated. These results are of home timeline tweets means tweets posted by people whom the registered user is following. Fig 4 shows the activity report of registered user.

Table 4: Polarity of home timeline tweets

id	user_id	Polarity	id_timeline_details
26995	7587032	Negative	569139627954741000
27009	22910295	Negative	569511413837737000
27023	7587032	Negative	569507822930677000
27027	14569869	Negative	569509161467285000
27032	14569869	Negative	569507348923977000
27033	14569869	Negative	569506871180173000
27039	21866939	Negative	569511791648251000
27196	612473	Negative	570161441443082000
27221	15995155	Negative	570765270656004000
27280	87818409	Negative	570819011581255000
27355	18413583	Negative	572673140825825000
27362	18413583	Negative	572672805860339000
27363	15063107	Negative	572670814983626000
27383	134758540	Negative	574848353084702000
27429	8443752	Negative	574867192472186000

The highlighted row in **Fig 4** indicates that users is social disturbing, on-click of that row navigate to page showing that particular user tweets and action to view user personal information (**Fig 5**) and to send message.

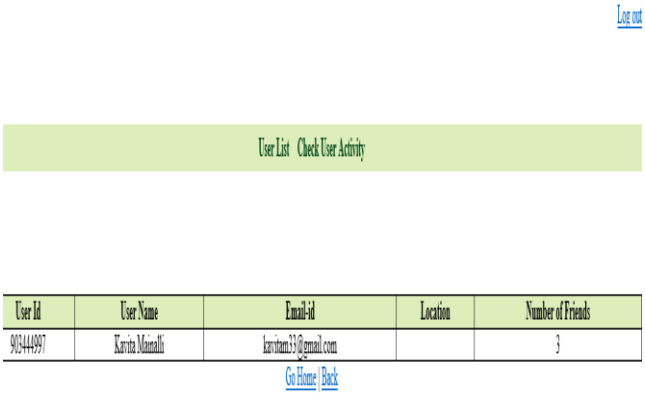

Fig 5: User Information

Conclusion and Future work

A new algorithm that identifies short text in tweets and replaces it with its meaning is developed in this work. The proposed work is focused on obtaining sentiment of tweets that shows positive or negative attitude of user towards society.

This research work is able to identify the social disturbing tweets and computing severity of such tweets. Based on severity and activity (strong negative or negative) it warns the user to stop such tweets. This work deviated from business oriented sentiment analysis to identify sentiment behind the tweets that may affect social peace.

Comparative analysis is done between StanfordNLP and Semantria. StanfordNLP provides more expected results; hence StanfordNLP is used for this research work. Scope exist to incorporate new mechanisms of sentiment analysis to improve performance.

References

[1] Scott A Brandt et.al – "Efficient Metadata Management in Large Distributed Storage Systems": Proceedings of 20th IEEE / 11th NASA Goddard Conference on Mass Storage Systems and Technologies 2003.

[2] Ainur Yessenalina, Yisong Yue and Claire Cardie : "Multi-level Structured Models for Document-level Sentiment Classification", Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

[3] Wojciech Gryc and Karo Moilanen : "Leveraging Textual Sentiment Analysis with Social Network Modelling" : Sentiment Analysis of Political Blogs in the 2008 U.S. Presidential Election. T2PP Workshop, 9-10 April 2010, Vrije Universities Amsterdam.

[4] Chenhao Tan, Lillian Lee, Jie Tang , Long Jiang, Ming Zhou, and Ping Li :" User-Level Sentiment Analysis Incorporating Social Networks ", conference conducted by ACM 2011.

[5] Oscar T¨ackstr¨om and Ryan McDonal: "Semi-supervised latent variable models for sentence-level sentiment analysis" 2011.

[6] Erik Tromp : " Multilingual Sentiment Analysis on Social Media": Master thesis July 2011.

[7] Aurangzeb khan, Baharum Baharudin : "Sentiment Classification by Sentence Level Semantic Orientation using SentiWordNet from Online Reviews and Blogs" : 2011

[8] Bing Liu- "Sentiment Analysis and Opinion Mining": Morgan & Claypool Publishers, May 2012.

[9] Hassan Saif : " Sentiment Analysis of Microblogs" Mining the New World , Technical Report KMI-12-2 March 2012.

[10] Mike Thelwall, Kevan Buckley – "Topic-Based Sentiment Analysis for the Social Web:The role of Mood and Issue-Related Words ": Journal of the American Society for Information Science and Technology 2012

[11] Anurag Mulkalwar, Kavita Kelkar : "Sentiment Analysis on Movie Reviews Based on Combined Approach" , International Journal of Science and Research (IJSR) 2012 .

[12] Diana Maynard et.al –"Multimodal Sentiment Analysis of Social Media", 2013.

[13] Anders Westling- "Sentiment Analysis of Microblog Posts from a Crisis Event using Machine Learning": 2013 Thesis, KTH Computer Science and communications.

[14] Duyu Tang et.al – "Learning Sentence Representation for Emotion Classification on Microblogs", Springer ,2013

[15] Xia Hu et.al –"Exploiting Social Relations for Sentiment Analysis in Microblogging " ACM , Rome, Feb, 2013

[16] M. Sakthivel , G. Hema Erik- "Sentiment Analysis Based Approaches for Understanding User Context in Web Content" : International Journal of Computer Science and Mobile Computing , July 2013.

[17] Infosys Briefings – "Bigdata Challenges and Opportunities" : 2013

[18] Gautham Vemuganti – "Metadata Management in BigData" : 2013

[19] Infomratica – "Metadata Management for Holistic Data Governance": 2013

[20] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts" Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank" 2013.

[21] Anurag Mulkalwar & Kavita Kelkar- "Sentence Level Sentiment Classification Using HMM with the Help of Part of Speech Tagging ", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) Oct 2014.