THE GOOD, THE BAD AND THE UGLY: OPINION MINING ANALYSIS ON USER TWEETS IN TWITTER

Joseph Philip[1], Ann Baby[2](Corresponding Author) and A.Kannammal[3]
[1,2]Rajagiri College of Social Sciences, Cochin, Kerala
[3]Coimbatore Institute of Technology, Coimbatore, Tamil Nadu

Abstract : Opinion mining evaluates people's opinions, sentiments, assessments, outlooks, and feelings from a written language. One of the most dynamic research areas in natural language processing, it is also extensively studied in fields of data mining like web mining and text mining. Opinion mining also finds its vast applications in management sciences and social sciences due to its importance to business and society asa whole. The emergent importance of opinion analysis parallelswithexpansion of socialmedia such as Facebook, Twitter, and social networks. This research suggests a stand-alone platform which enables users to perform opinion analysis which makes use of Twitter API to gather information aboutthe tweets and its users. The proposed model is used to analyse whether the opinion in a text is neutral, positive ornegative. As users tend to post frank and blunt comments in twitter, it has emerged as a platform to check people's feedback.The inclusion of hashtags to attract social attention also is a trigger for opinion mining. Twitter Search and Streaming APIs with the Stanford NLPlibrary are used to opine the data collected. The pre-processed data passes through the opinion generation phase where opinion classes are defined for polarity. Words are labelled with opinions, and the general opinion of a tweet is thus analysed, in this research paper.

IndexTerms **- Twitter, Opinion Mining, Opinion Classes.**

I. INTRODUCTION

Opinion Mining, is the method of computationally classifying and labelling opinions(feelings/sentiments) articulated in a portion of text, to conclude aboutthe writer's attitude towards an idea is, negative, positive or neutral using the concept of computational linguistics[1]. The idea could be a topic or product on which the user can opine freely. The customer's opinion for a product, also known as Voice of Customer (VOC) is crucial for organizations [2]. VOC is used widely to analyse customer's needs, infer his wants and to obtain timely feedback from the customers. Nowadays, VOC is widely considered as a market exploration and research procedure to produce a thorough and comprehensive picture of customer wants and needs, organized into a hierarchical arrangement, and then ranked in terms of relative prominence and satisfaction with current replacements.

1.1 Opinion Analysis

Opinion analysis is the process of contextually mining text to identify and categorize the subjective opinions expressed by the users and writers[3]. It is widely used to decide whether the user's/writer's attitude towards a particular topic or product, etc. isnegative, positive or neutral[4]. It is also used by businesses to help them apprehend the social sentimentality of their product, brand, or services while observingand monitoring online conversations[5].

1.2 Mining and Analysis of Twitter Data

Twitter, the gold mine of data is different from other social platforms. The tweets of almost all users are public and retrievable. The asymmetric network infrastructure of "friends" and "followers" in Twitter permits a user to include tweets from accounts whose content the user finds current and interesting. The incredibly high number of users in Twitter adds to the fact that it is widely used as a marketing tool. It is also used as a transport layer for third-party messaging services [6]. Thus Twitter has a widespread collection of APIs, to build customised networks.

Web-blogs are used to construct bodies for opinions analysis and emotion icons are treated to be mood indicators [7]. Conditional Random Fields (CRF)and Support vector Machines (SVM) are used to classify opinions at sentence level. Various strategies are analysed to define the overall opinion of a document. The winning structure is thus defined by considering the opinion of the last sentence of the document.

1.3 Opinion mining of Twitter Data

Emoticons like ":-)" and ":- ("are used for opinion mining to form the training set for classification. In such researches, texts containing emoticons are collected from the public social domain and the dataset is divided into "positive" (texts with happy emoticons) and "negative" (texts with angry or sad emoticons) samples. SVM (Support Vector Machines) and Naive Bayes (Emoticon strained classifiers) produce 70% of an accuracy on the test dataset[8]. Twitter is also used to elicit training data to perform opinion searches. Corpora are created with emoticons to obtain "negative" and "positive" samples, and then classifiers are used. Naives Bayes classification produced a mutual information measure for the feature selection in Twitter with up to 81% of accuracy on the test dataset, though bad performancewith three classes ("negative", "positive" and "neutral") were observed [9]. Natural Language opinions are regularlyarticulated in meticulous and variedmethods, giving queries which are challenging to give details by simple text processing techniques. However, approaches like applying n-grams, Part-of-speech tagging have been used efficiently to find the twitter opinion using machine learning techniques and other methodologies[10]. There are also methods which employ the literal meaning of the word to classify into negative or positive classes, irrespective of the target while performing opinion analysis. Thus understanding the target and the relative terms of the opinion analysis is important[11].

II. THE PROPOSED MODEL FOR OPINION ANALYSIS IN TWITTER

In opinion analysis, the mood of a comment or tweet can be quantified by the number of negative and positive words which occur in the comment or tweet. A score can be generated by subtracting the number of negative counts from the positive counts[12]. At present, information can be gathered through individual interviews, online questionnaires, contextual enquiries, ethnographic methods etc. All these methods involve a sequence of structured in-depth interviews, which give focus to the experience of customers about the products and alternatives they are currently endorsing. Theneeds statements are thenmined andstructured into a more usable hierarchy, and laterranked by the customers[13]. The process is a tedious one where analysts need to examine the entire data, extract the opinion about the statement and then determine the statement's opinion. The proposed system is a stand-alone web application which enables the user to perform the opinion analysis. In this research, information available through the Twitter API is utilized togather information about the tweets and their users. As Twitter has a restriction where tweets are to be less than 140 characters in length,comments of users usually have a tendency to be frank, straightforward and blunt. In addition, due to the large influence of Twitter people are lured to include hashtags in their tweets so that h they can gain social attraction and prominence.Thus, Twitter has become a great platform to study people's criticism and feedback[14]. This research which focusses on the "opinion-filled, sentiment-filled" Twitteruses Twitter Search and Streaming to retrieve the tweets and Stanford NLP library to detect the opinions. Results are stored onto a database on a real-time basis which can be retrieved later. The schematic workflow of the proposed opinion mining technique is described in Figure 1.

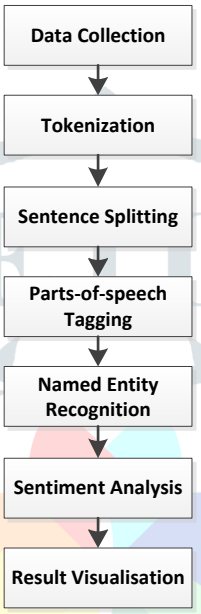

Figure 1 – Schematic workflow for opinion mining in Twitter

In the initial phase of data collection,users register through a registration module where the system permits users to enrol themselves with their personal details along with their user name, password and with a pair of Twitter authentication tokens foraccessing the Twitter APIs. Users with the help of their customisable dashboard, users can either upload a text document or can manually enter the desired texts in the provided user interface to perform the opinion analysis and the results can be stored to database if required for further retrieval. Data is pre-processed and cleaned for stopwords removal, punctuation removal, stemming etc. Tokenization splits the tweets into relevant tokens. As tweets may containseveral opinions about the data which are expressed differently by various users, they need to be tokenized. The twitterdataset used in this research work is already labelled into two classes of negative and positive separation. This enables theopinion analysis of the data becomes easy to observe the effect of various features[4]. Stop words removal has been omitted in this research as it decreases the performance, as it depends entirely on the tokenized data. Sentence splitting is performed to together with tokenization, as the opinions of each sentence need to be determined. Twitter user-defined hashtags in tweets are used for classification of opinion type using punctuations, single words, n-grams and patterns as different feature types, which are then combined into a feature vector for opinion classification. SVM strategy is used to assign opinion labels by constructing a feature vector for each example in the training and test set [15]. Parts of speech like nouns, nouns, proper nouns and Named Entity Recognitions (NER), which are good pointers of biased opinions are used to generate patterns, as they are good indicators of opinions. Users can fetch the social media subjects related to hashtags or profile content to perform the opinion analysis. The content is fetched via Twitter's API with the help of the authentication token which is provided on the sign up process. Contextual mining is also performed in this phase in cases where users tweet continuously about a product (noun). In the NERphase, Java implementation of the Stanford NER is adopted [16]. In NER, word sequences in a test are labelled as either names of persons, companies, places, things etc.Feature extraction is performed in this stage where major recognizers for the English language are classified into three classes, namely PERSON, ORGANIZATION and LOCATION. Training for this is performed on the CoNLL 2003 English training data. In the below example for the English sentence "I am from Delhi", the LOCATION class is tagged.

Eg: I am from Delhi
ORGANIZATION
LOCATION
PERSON

The complexmodel of opinion mining analysis is omitted in this research. The simple model is followed where only negative, positive and neutral opinions are mined and classified. In the proposed model, opinions are classified into five (0-4) classes. Table

1 depicts the classification of the classes based on their polarity.Natural language processing is used to tag each word with anopinion and the overall score or opinion that the statement would get, depends on the underlyingword opinion.

Table 1. Opinion Classes

Class	Opinion
0	Very Negative
1	Negative
2	Neutral
3	Positive
4	Very Positive

Opinion trees are constructed from the tweets based on the maximum class obtained at each node of the tree. Figure 2 depicts the opinion tree for the English text "SC likely to hear Ayodhya appeals from July 25" [17]. Figure 3 depicts the graphical opinion of the final opinion node, where very negative shows a score of 6, negative 36, neutral 35, positive 20 and very positive 4. Thus, the above sentiment shows a final "Negative" opinion.

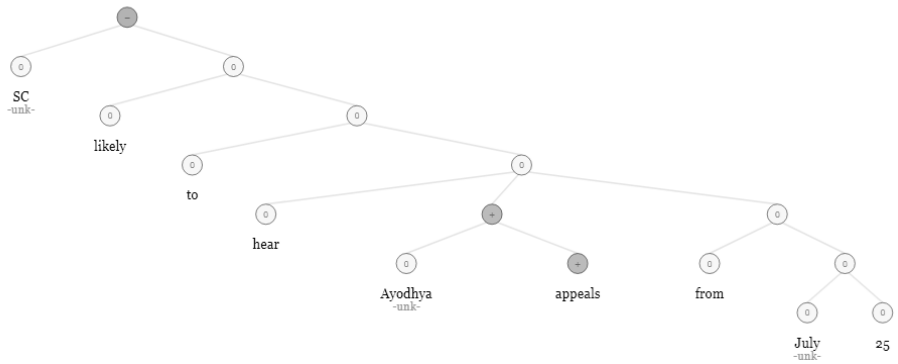

Figure 2 – Opinion Tree for "SC likely to hear Ayodhya appeals from July 25" [17]

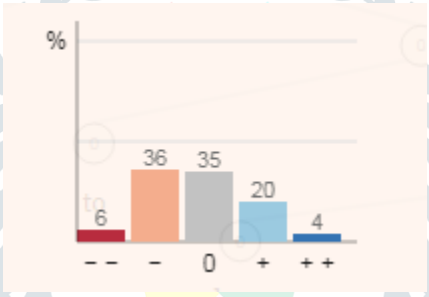

Figure 3 – Graphical Opinion of the Final Opinion Node.

In contrast with the Stanford Core NLP Package which takes input from the user, the proposed system takes in input from real-time tweets. For each sentence, theopinion value is marked with legends. Metadata of the tweet which includes individual user details, all users involved in the process, user location, NERs, nouns, verbs, date/time of tweet, and all other relevant details of the tweet are highlighted[18]. The system is integrated with a Representational State Transfer (REST)web server so that the user can perform the opinion analysis over the Internet, irrespective of the network. Thus, guest users can also avail the analysis facility to overcome the desktop, to offer as a service

III. FINDINGS

the Colour Legends are attached to each opinion in the proposed model. The colour coding schemes as represented in figure 4 are followed for the opinion analysis.

Opinion	Legend
Very Negative	
Negative	
Neutral	
Positive	
Very Positive	

Figure 4 – Colour Legends for Opinions

Text Opinion gathering does a simple opinion analysis based on the colour legends and produces the opinion for each line/word in the given text.The opinions could range from very negative to very positive. Figure 5 shows the opinions for a two- paragraph text from the 2014 Nobel Peace-Laureate Malala Yosusafzai's[19] speech to the United Nations, which has the following text: "I don't know where to begin my speech. I don't know what people would be expecting me to say. But first of all, thank you to God for whom we all are equal and thank you to every person who has prayed for my fast recovery and a new life. I cannot believe how much love people have shown me. I have received thousands of good wish cards and gifts from all over the world. Thank you to all of them. Thank you to the children whose innocent words encouraged me. Thank you to my elders whose prayers strengthened me.

I would like to thank my nurses, doctors and all of the staff of the hospitals in Pakistan and the UK and the UAE government who have helped me get better and recover my strength. I fully support Mr Ban Ki-moon the Secretary-General in his Global Education First Initiative and the work of the UN Special Envoy Mr Gordon Brown. And I thank them both for the leadership they continue to give. They continue to inspire all of us to action." [20]

I don't know where to begin my speech.

I don't know what people would be expecting me to say.

But first of all, thank you to God for whom we all are equal and thank you to every person who has prayed for my fast recovery and a new life.

I cannot believe how much love people have shown me.

I have received thousands of good wish cards and gifts from all over the world.

Thank you to all of them.

Thank you to the children whose innocent words encouraged me.

Thank you to my elders whose prayers strengthened me.

I would like to thank my nurses, doctors and all of the staff of the hospitals in Pakistan and the UK and the UAE government who have helped me get better

I fully support Mr Ban Ki-moon the Secretary-General in his Global Education First Initiative and the work of the UN Special Envoy Mr Gordon Brown.

And I thank them both for the leadership they continue to give.

They continue to inspire all of us to action.

Figure 5 – Opinions for Malala's speech

The overall opinion analysis is depicted in Figure 6, which shows a result of "Slightly Negative".

Total very postive opinion value: 0
Total postive opinion value:6
Total neutral opinion value:0
Total negative opinion value:6
Total very negative opinion value:0

Final opinion value:1
Final opinion: Slightly negative

Figure 6 – Analysis of Malala's speech

In the proposed model, Twitter Keyword searching gathers the tweets from twitter which have the mentioned keywords, and opine it based on the context of the text[21]. The gathered tweets are scrapped off their user credentials, location details, and timestamps in the pre-processing stage. Once the opinion analysis is performed, the tweets are colour classified based on the colour coding scheme adopted for the study. Figure 5 shows the opinion analysis for the keyword "World cup 2019". Latest tweets (tweets occurred in the last seven days) are gathered through the twitter-developer API and the red tweets show negative opinions, and the green tweets show positive opinions. Figure 7 depicts the twitter keyword searching for "World cup 2019".

I am too saying that they deserved to the final but not the wordcup.

So it had to be England.

New champions @ECB_cricket, deserved champions.

Very bad icc rules.

Wordcup final dicide by boundary.

Just not fair.

#ICCRules @ICC @cricketworldcup @ICCMediaComms @englandcricket @cricbuzz @BLACKCAPS @BBCWorld

England 241 lost 10 wickets
New Zealand 241 lost 8 wickets ...
But England won the match....

Bullshit win for England.....!

Wordcup final will be remembered as a tie Match..... ....!!!

The best cricket's wordcup final, but luckier team winning based on boundaries hit is bit anti climax.

#CWC19 #CWC19Final

Congrats Eng, tough tough luck NZ.

England won the WordCup but NewZealand won the Hearts <3 Congrats to both of them!

#Great_Match
#Final
#CricketWorldCup2019

Figure 7 – Twitter keyword Search for "World Cup 2019"

The proposed model provides an extensive tweet analysis which consists of details of the tweets for a particular keyword, which include tweet time, user details, tweet URL, number of followers, description of the username, display picture(hyperlink), status count(if retweeted), geo-location and account verification details. The same is depicted in Figure 8, which shows the tweet details for the text "World cup 2019".

Tweet	Tweet tweeted time	URL	Username	Email	Description	Display picture	NO. of Follow
Every one is saying t...	Tue Jun 11 20:35:04 I...	<a href="...	Lone Mudasir	<INFO:privac...	https://t.co/KXaGGX...	hidden	0
Very bad icc rules. W...	Sun Jun 05 22:39:20 ...	<a href="...	Mohit Chaubey	<INFO:privac...	Have you ever felt s... makes sense? Well t... now...I feel like I'm f...	hidden	9
England 241 lost 10 ... New Zealand 241 los...	Sun Oct 18 14:50:32 ...	<a href="...	Maeraj Manzoor	<INFO:privac...	Maybe one day you ...	http://abs.twimg.co...	1856
The best cricket's wo...	Tue Aug 05 04:08:18...	<a href="...	Adi Vemuru	<INFO:privac...	Cloud Solutions Arc...	http://abs.twimg.co...	1120
England won the Wo... #Great_Match #Final	Thu Nov 29 12:54:43...	<a href="...	Abul Bashar Arman	<INFO:privac...	A great Writer, Poet ... An Entrepreneur and... Former RJ (Radio Joc...	http://abs.twimg.co...	481

Figure 8 –Tweet Analysis for "World Cup 2019"

Named Entity Recognition for the various tweets are filtered on the basis of parts of speech and the nouns/subjects and detailed in the proposed model as in Figure 9.

Named Entity Recognition	@users	Display Name	Geo Location
one	Lone Mudasir	mudasirkj	<unknown>
Newzeland	Mohit Chaubey	Mohit_Chaubey	Rewa, India
#WC19	Maeraj Manzoor	maeraj_manzoor	Srinagar Jammu And Kash
wordcup	Adi Vemuru	vemuruadi	Singapore
England	Abul Bashar Arman	AbulBasharArman	islamabad, pakistan
champions			
@ECB_cricket			
champions			
icc			
rules			
dicide			
boundary			
#ICCRules			
@ICCMediaComms			
@BLACKCAPS			
@BBCWorld			
England			
wickets			
New			

Figure 9 – NER for "World Cup 2019"

Theproposed models is also provided as a service through the REST API server. Keyword searching is enabled through the API call:/myapp/sentiments/text?text= <keyword to be searched>. Twitter searching is enabled through the API call: /myapp/sentiments?searchKeywords= <tweet keyword to be searched>. Figure 10 shows the analysis of searching "Microsoft", where the JSON output from the REST API is rendered to a HTML format.

Figure 10 – Opinion Analysis for "Microsoft" through the REST server.

I. SCOPE FOR FUTURE WORK

The proposed system performs opinion mining of user tweets on the social media. Though optimal accuracy is achieved, the system needs further enhancement to deal with sarcasm(E.g.: Brutus is an honourable man). Thwarted comments and frustrations in tweets can be a further area of study using complex real-time models. Tweets including a negative and a positive word (E.g.: Not good) could lead to neutrality, though the final answer rests on the data in the tweet. There are also cases of faulty acceptances, which can be a further area of study using advanced natural language processing tools.

IV. CONCLUSION

This research can be plugged into product based enquiry tools and online review platforms to gather general opinions from the public tweets. Opinion analysis is widely used in applications that use reviews from websites which include product reviews, comments about a particular service, feedback on political issues etc. It is also now included as a sub-component in various recommender systems. It also has wide applications in business intelligence where opinions are extracted from online reviews to improve the products and thus increase customer satisfaction. It is also finding applications in sports, medical and smart home deployment areas. Opinion analysis clubbed together with Internet of Things (IoT) can be employed for further research. Opinion analysis will also play a major role in real time trend prediction analysis.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1, pp. 1-135, 2008.

[2] C. Aguwa, M. H. Olya and L. Monplaisir, "Modeling of fuzzy-based voice of customer for business decision analytics.," *Knowledge-Based Systems*, vol. 125, no. 1, pp. 136-145, 2017.

[3] F. Benamara, M. Taboada and Y. Mathieu, "Evaluative language beyond bags of words: Linguistic insights and computational applications," *Computational Linguistics*, vol. 43, no. 1, pp. 201-264, 2017.

[4] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of techniques.," vol. arXiv preprint arXiv:1601.06971., 2016.

[5] S. Mukherjee and P. Bhattacharyya, "Sentiment analysis: A literature survey.," vol. arXiv preprint arXiv:1304.4520., 2013.

[6] M. A. Russell, Mining the Social Web, Publisher: O'Reilly Media, Inc., Februray 2011.

[7] Y. Changhua, H.-Y. L. Kevin and C. Hsin-Hsi, "Emotion classification using web blog corpora.In," in *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,IEEE Computer Society.*, Washington, DC, USA. , 2007.

[8] R. Jonathon, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," *The Association for Computer Linguistics*, vol. 3, no. 1, pp. 43-48, 2005.

[9] A. Go, L. Huang and R. Bhayani, "Twitter sentiment analysis," Final Projects from CS224N for Spring 2008/2009., 2009.

[10] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," *43rd annual meeting on association for computational linguistics*, vol. 1, no. 2, pp. 115-124, 2005.

[11] B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," *Fourth International AAAI Conference on Weblogs and Social Media.*, 2010.

[12] S. Mukherjee and P. Bhattacharyya, "Sentiment analysis in twitter with lightweight discourse analysis," in *COLING 2012 (pp. 1847-1864).*, 2012 December.

[13] O. Turel, A. Serenko and P. Giles, "Integrating technology addiction and use: An empirical investigation of online auction users," *MIS Quarterly*, vol. 35, no. 4, pp. 1043-1062, 2011.

[14] E. Lahuerta-Otero and R. Cordero-Gutiérrez, "Looking for the perfect tweet. The use of data mining techniques to find influencers on Twitter," *Computers in Human Behavior*, vol. 64, pp. 575-583, 2016.

[15] P. Chikersal, S. Poria, E. Cambria, A. Gelbukh and C. E. Siong, "Modelling public sentiment in Twitter: using linguistic patterns to enhance supervised learning," *International Conference on Intelligent Text Processing and Computational Linguistics, Springer*, pp. 49-65, 2015.

[16] "Stanford University," [Online]. Available: https://nlp.stanford.edu/software/CRF-NER.html. [Accessed 8 6 2019].

[17] "The Hindu," [Online]. Available: https://www.thehindu.com/todays-paper/sc-likely-to-hear-ayodhya-appeals-from-july-25/article28393815.ece. [Accessed 12 7 2019].

[18] F. H. Khan, U. Qamar and S. Bashir, "A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet.," *Knowledge and information Systems,* vol. 51, no. 3, pp. 851-872, 2017.

[19] "Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Malala_Yousafzai. [Accessed 16 8 2019].

[20] "United Nations," [Online]. Available: https://www.un.org/News/dh/infocus/malala_speach.pdf. [Accessed 16 8 2019].

[21] S. Poria, H. Peng, A. Hussain, N. Howard and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing,* vol. 261, pp. 217-234, 2017.