IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Survey on Tweet Segmentation and Sentiment Analysis

S.S. Ansari[1*], T. Diwan[2]

[1*] Dept. of CSE, Shri Ramdeobaba College Of Engineering and Management, Nagpur,India
[2]Dept. of CSE, Shri Ramdeobaba College Of Engineering and Management, Nagpur,India

[*]*Corresponding Author: sharfuddinsr@rknec.edu*

Available online at: www.ijcseonline.org

Received: 26/Dec/2017, Revised: 30/Dec/2017, Accepted: 22/Jan/2018, Published: 31/Jan/2018

***Abstract*—** With the explosive growth of user generated messages, twitter has become a social site where millions of users can exchange their opinions. Sentiment analysis on twitter data plays an important role in finding public opinions which have provided an economical and effective way timely, which is very useful for decision making in various domains. A company can take the public opinion in tweets to obtain user review towards its products where a politician can adjust his position with respect to the opinion change of the public. There have been a large number of research studies and industrial applications in the area of public sentiment tracking and modeling. Millions of users give their opinions on Twitter, making it a valuable platform for tracking and analyzing public sentiment. Such tracking and analysis can provide critical information for decision making in various domains. So, it has attracted attention in both academic and industry. Previous researches showed that the tweet was classified appropriately only if the tweet would contain the exact same label as the training set. But this approach fails when the tweet contains a synonym or a variant of the label instead of the exact same label.

***Keywords*—** Classifier, Opinion Mining, Lexicon, Sentiment Analysis, Twitter

I. INTRODUCTION

Microblogging site such as Twitter has attracted millions of users to share information or viewpoints resulting in a large number of data produced every day. The information given on social media is delivered to every person within some fraction of a second. Twitter or Facebook is the platform which is being widely used for postings and it also gives the individual person to express every emotion on such social platform.
Tweet Segmentation is splitting a tweet into meaningful segment is called tweet segmentation.

For example: I said to expend your time in out door games.

This example (tweet) is separated into five segments as shown below. Semantically input segments "Expend your time" and "outdoor games" are sealed.
(I said) | (to) | (expend your time) | (in) | (outdoor games)

In this paper they focus on the task of tweet segmentation which splits a tweet into a sequence of consecutive n-grams, each of which called a segment. Named Entity Recognition (NER) is the subtask of information extraction that classify named entities such as the name of persons, organizations and locations in the text. They proposed a generic tweet segmentation framework, i.e. Hybridseg, which learns from both global context and local context and evaluate two segments-based Named Entity Recognition (NER) algorithms which are unsupervised in nature and the inputs are the tweet segment [1]. It defines the framework of segmentation of tweets by collection form, called as Hybridseg, and it summarizes the tweets by using a Document summarization algorithm [5], [21].

It describes the segment-base event detection system for tweets, called Twevent, in which first detect bursty tweets and perform segmentation that is splitting a tweet into segment as event segment and then calculate the frequency of segment based on predefined fixed time window (e.g. One day or One hour) and then performs clustering to group event-related segment [6]. In this system they focus on segmentation which splitting a tweet into segment, to achieve high quality of tweet segmentation they propose a framework, named FRAppE (Facebook's Rigorous Application Evaluator) it also detects the misuse in social media, and this system provides an automatic warning before uploading harmful messages [7]. In this they analyzed social media data (i.e. Twitter data, because twitter is most commonly used social media nowadays), and the most common use of social media is to mine customer sentiment, so they implement KNN (K-nearest neighbor) approach to eliminate rumor based tweets or negative tweets with improved accuracy rates [9]. Twitter is most popular platform and many user share different types of URL which may be harmful for another user and malicious URL a.k.a malicious website is a common and very serious threats of cyber security, It describes the generality of malicious URL detectors by using machine learning techniques [17]. Tweet classification helps to maintain data region wised. With the help of a data mining algorithm the data must easily maintain and easy to access. In this paper they proposed a tweet

classification, data mining algorithm is used for classification of tweets and k-means algorithm is also used for event detection [22].

Rest of the paper is organized as follows, Section I contains the introduction of tweet segmentation and some basic techniques used in the procedure. Section II contain the related work of how microblog tweet segmentation and sentiment analysis is done in different papers using different techniques and also the description of techniques are also mentioned which were used in them. Section III contain some limitiation from the techniques present so that to implement technique which can overcome this limitations. Section IV contain the results and discussion followed by future scope in section V.

II. RELATED WORK

A. Linguistic Features

It explores the connection between the web mining categories and the related age not a paradigm. For this survey, focus is on the representation issues, on the process, on the learning algorithm and on the application of the recent works as the criteria. Exploit's the local linguistic features in a collective manner by using the existing NER tools. The recognized names entities with high confidence positively enhance the performance of tweet segmentation. For named entity recognition both supervised and unsupervised approaches have been used and it is based on CRF model. Tweet segmentation is conceptually similar to Chinese word segmentation, In Chinese text is a continuous sequence of characters. State-of-the-art CSW methods are developed by using supervised learning techniques like CRF model and perceptron learning [1]. It describes the research done for the information retrieval giving an IR view of the unstructured documents. Also, the information retrieval view of semi-structured documents is discussed. The database view of the web content has been explained in detail which mainly tries to model the data on the web and to integrate them so that sophisticated queries other than keywords based search could be performed [2]. On web mining it describes the importance of filtering to get knowledge about the actual usage of a website by the calls made from the robot [3].

B. Data patterns

This is patterned between different web pages and creates more customized and accessible web pages to users, which in turn creates more traffic and trade to the website. Also, some common methods to find and eliminate the web usage made from robots while keeping browsing data made from human users intact are addressed. Web mining is viewed as seen to consist of three major parts: collecting the data, preprocessing the data and extracting and analyzing patterns in the data [4].

C. Mining and Visualization

Worked on a survey of the existing techniques of web mining and the issues related to it. It primarily illustrates the summary of various techniques of web mining approached from the various angles like feature extraction, transformation, representation and data mining techniques in various application domains. The survey on data mining technique is made with Clustering, classification, sequence pattern mining, association rule mining and visualization. It also gives the overview and some important research issues related to web mining. It describes the process of web cleaning, which is needed to remove noise and correct inconsistencies in the data. Discussed the overview of different web content mining tools. Extracting useful information from the web documents is the one of the process of Web content mining. For downloading the essential information that one would require the content mining tools help with the flood of information and data on the Web [5]. Made a comprehensive study of the various web content mining techniques, tools & algorithms. In previous work, document pivot methods are used for topic detection and tracking. But it causes cluster fragmentation problems because in this, the topic is represented by a cluster of documents. First, they used hashtags method, then it is converted into tf-idf method, but the important drawback of the method is the need for manual annotation of training and test samples. Also sensitivity to noise is known problem in a document pivot method, so after that the latent sensitive method is used, but such a solution is problematic when the nearest neighbor are not very similar. In previous work, feature pivot methods are used for first story detection. They used latent document allocation, tf-idf graph based approaches, but it causes misleading term correlation. Misleading term means in input searching for one term and in output it gives many different term related result which is not the requirement [6]. Proposed an algorithm for structure data extraction [7]. It described three recent extraction systems that can be operated on the entire web [8].

D. Classifier

Analyzed the nature and distribution of structured data on the web discussed various approaches and applications of text mining [10]. Primarily focused on enhancing the relationship between text mining and data mining [11]. It made a survey about different text mining techniques and applications [12]. Discussed the study of the problem of determining opinion of users, which are expressed on product features in reviews, it may be positive, negative or neutral [13]. Describes supervised opinion mining techniques on online user reviews. The proposed algorithm in detects sentiments on movie user reviews, based on NB classifier [14]. A growing number of recent studies have been focused on the opinion

mining, it proposed a modified algorithm which mainly focuses on the experts opinions, so that the readers could understand the content easily [15]. This made a survey of work done by different researchers on sentiment analysis and opinion mining techniques [16]. In this, they analyze social media data. The most common use of social media is to identify the users review on a product to support marketing and customer service activities and also get reviews on politics and so on, the user review may be positive, negative or neutral. And then take twitter big data to predict named entity. Twitter is widely being used for posting, by considering wide use of Twitter as the source of information, it is challenging for reaching an interesting tweet among a bunch of tweets for a user. In this work, it is aimed to reduce the Twitter user's effort to access to the tweet carrying the information of interest. To achieve the goal, Hybridseg is generated via named entities extracted from user's followers' and user's own posts [17]. Extend our approach to analyze short text in tweets and rumor based tweets. So they implement K-nearest neighbor classifier (K-NN) approach to eliminate rumor based tweets or negative tweet from twitter with improved accuracy rates and to identify the rumor with high level security, we can implement real time tweet environments in future [18].

E. Candidate segment

Twitter has attracted millions of users to share information or up-to-date information, resulting in a large number of data produced every day. Even though, at various applications of Natural Language Processing and Information Retrieval go through rigorously from an erroneous and tiny nature of tweets. To implement a framework in support of segmentation of tweet by collection form, called as HybridSeg. During tweet separating with trivial segments, surroundings information is preserved and simply takes out by the downstream application. Hybridseg glance for top segmentation of a tweet through increasing stickiness scores of its candidate segment[19].The stickiness score is explained the possibility of a segment is expressed in English (global context and local context). Finally, we advise and assess two models to acquire with local context by concerning the term dependency in a collection of tweets, in the same way. Testing on two tweet data sets gives you an idea about the tweet segmentation superiority is considerably enhanced by global and local contexts evaluate by use of global context simply. Assessment and relationship, we demonstrate that additional correctness is accomplished in Named Entity Recognition by part-of-speech (POS) tagging of placing segment-based [20].

III. LIMITATION

Millions of users share opinions on different aspects of life every day in popular websites such as Twitter, Tumblr and Facebook. Spurred by this growth, companies and media organization are increasingly seeking ways to mine these social media for information about what people think about their companies and products. Political parties may be interested to know if people support their program or not. Social organizations may ask people's opinion on current debates. All this information can be obtained from micro blogging services, as their users post their opinions on many aspects of their life regularly. In this work, we present a method which performs classification of tweet sentiment in Twitter. Need to provide end to end system which can determine the sentiment of a tweet at two levels- phrase level and message level. The features of tweets to build the classifier which achieves accuracy at phrase level and at the message level. The tweet segmentation quality is significantly improved by learning every international and native context compared with victimization international context alone. Event detection from Twitter stream is challenging for at least three reasons noisy content, diverse and fast changing topics, and large data volume. Effectiveness of utilizing more features from tweets like retweet rate and hash tags in Twevent. Another important task is to investigate the effectiveness of Twevent. The previous work in that the named entity extraction (NEE) and linking for the tweets it is the hybrid approach. The named entity extraction is for locating phrases in the text that represent the names of persons. The approaches are that named entity generation and linking then its filtering.

IV. RESULTS AND CONCLUSION

They proposed Hybridseg framework for tweet segmentation, Hybridseg learns from both local and global contexts and also able to learn from pseudo feedback. It mainly focused on the evaluation of the Named Entity Recognition (NER) and reports the NER accuracy of the five methods which are unsupervised method. In which $Unigram_{pos}$ is the worst performer among all method and $Hybridseg_{pos}$ achieves the best result , they conclude that in future we can improve the quality of segmentation and explore the effectiveness of tweet[1]. They improve the algorithm, enriching the training set of examples, on the way, with examples classified as strong positive or negative, by an established score of classification. They try to highlight the main aspects on which opinions are expressed and to extract opinions based on aspects identification. They compare Twevent with the other two methods, i.e. EDCoW and $Twevent_u$ and show that Twevent is the best method which yields the best precision of 86.1%, which is significantly larger as compare to the precision achieved by EDCoW and $Twevent_u$ [6].

In this work they observed over a nine month on malicious Facebook apps and they developed FRAppe apps for detecting malicious Facebook applications, it is an accurate classifier [7]. In the previous approach the classifier

cannot classify the tweet accurately, because the sentiment, word used in the sample test and in the training test are not same even if it have the same meaning. In this approach they solve this problem by collecting or gathering most frequently used sentiment, word and their possible synonyms having the same meaning. The second problem is the flexibility of using different variations of sentiment word. In this approach they solve this problem by allowing users to use variations of sentiment words and it increases the accuracy of the classifier and provide flexibility [14]. This work mainly focused on tweet classification and the data mining algorithm properly handled the data and classified it region wised which helps to improve the accuracy and efficiency of the tweets. It provides the current event detection, which is also useful for the traffic analysis [22].

V. FUTURE SCOPE

Facebook will benefit from our recommendations for reducing the menace of hackers on their platform. They try to determine, in a review, those sentences which do not express opinions, or determine opinions about the film or the film actors and identify opinions addressed strictly on these items. For future work we can improve the segmentation quality and calculate the effectiveness of tweets and also improve the graphical analysis. And we can protect data from spam and hence the tweets are secured in nature.

VI. REFERENCES

[1] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, Member IEEE, "*Tweet Segmentation And Its Application To Named Entity Recognition",* IEEE Transactions on knowledge and data engineering ,Vol.27,No.2, pages. 558-570 , 2015.

[2] Ana C. E. S. Lima, Lean Dran. Castro, "*Development of a Novel Algorithm for Sentiment Analysis Based on Adverb Adjective Noun Combinations*", IEEE, 2012.

[3] R. Varghese, "*A Survey on Sentiment Analysis and Opinion Mining*", IJRET, 2013.

[4] Ana C. E. S. Lima, Lean Dran.Castro "*Automatic sentiment Analysis of Twitter Message*", IEEE, 2012.

[5] Magar Ranjeet, Bhoge Swapnil, "*A Survey on Tweet Segmentation and its Application to Named Entity Recognition*", IJIRCCE, 2016.

[6] Chenliang Li, Aixin Sun, Anwitaman Datta " *Twevent: Segment-based Event Detection from Tweets*", ACM, 2012.

[7] M. Ganga, S. Aanjan Kumar,"*Segmenting and Detecting Malicious Tweets and Harmful Entity Recognition*" , IJIRCCE, 2016.

[8] Hu X. Tang, "*The How, When and Why of Sentiment Analysis*", IJCTA, 2013.

[9] R. Gomathi, M. Rajkumar, " *Tweet Segmentation And Classification For Rumor Identification using KNN Approach*", IJCRME, 2016.

[10] Shachi H. Kumar, University of California, Santa Cruz Computer Science, "*Twitter Sentiment Analysis CMPS 242 Project Report*".

[11] Ian H. Witten, "*Text Mining*", pp. 1-23.

[12]Vishal Gupta , Gurpreet S. Lehal, "*A Survey of Text Mining Techniques and Applications*", JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, Vol.1, 2012.

[13] Xiaowen Ding, Bing Liu, Philip S. Yu, "*A Holistic Lexicon Based Approach to Opinion Mining* ", ACM, 2008.

[14] Ion Smeureanu, Cristian Bucur, "*Applying Supervised Opinion Mining Techniques on Online User Reviews*", Informatics, Economics, Vol .16, 2012.

[15] K. Nathiya, Dr. N. K. Sakthivel, "*Development of an Enhanced Efficient Parallel Opinion Mining for Predicting the Performance of Various Products*", International Journal of Innovative Research in Computer and Communication Engineering, Vol.1, 2013.

[16] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, "*OPINION MINING AND ANALYSIS: A SURVEY*", International Journal on Natural Language Computing, Vol.2, 2013.

[17] Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi " *Malicious URL Detection using Machine Learning: A Survey*", IEEE, 2017.

[18] V. Gayathri, A.E. Narayanan "*Tweet Segmentation And Classification For Rumor Identification Using KNN Approach*", Indian J. SCI. Res. 14 (1): 102-108, 2016.

[19] S. Kukku S, Reshma Reghu and Gaina K.G, "*Tweet Segmentation and its Application Using RandomWalk And Part-of-speech Methods*", I J C T, pp. 7497-7501 , 2016.

[20] MandhalaVinoothna, "*Segmentation of Trust Worthy Based Secure Data*", International Journal of Big Data Security Intelligence Vol.2, No.2, pp. 23-28 ,2015.

[21] Chetan Chavan, Ranjeet Singh, "*Summarization of Tweets and Named Entity Recognition from Tweet Segmentation",* International Conference on Automatic Control and Dynamic Optimization (ICACDOT),International Institute of Information Technology, pages. 66-71 , 2016.

[22] Sonam Meshram, Hirendra Hajare, "*Tweet Segmentation and Enhancement of Tweets*", International Journal of Science and Research (IJSR), Volume.5 Issue.5, pages. 577-579, 2016.

[23] Prof. Vikas Balasaheb Burgute, Prof. A. K. Gupta, "*Named Entity Recognition using Tweet Segmentation* ", International Research Journal of Engineering and Technology (IRJET), Vol: 4, Issue: 7, pages. 1068-1075, 2017.

Authors Profile

Miss.Shabistan Ruhi Ansari has done Bachelor of engineering in computer Science and engineering from Rashtrasant Tukdoji Maharaj Nagpur University,India in 2015 and currently pursuing P.G in computer science and engineering from Ramdeobaba college of engineering and management,Nagpur,India.And research work focus on Data Mining,Information Retrieval and machine learning.

Dr.Tausif Diwan received M.Tech. And Ph.D. in Computer Science and Engineering from VNIT college, Nagpur, India in 2011 and 2017 respectively. Since June 2012, he has been with the Department of Computer Science and Engineering, RCOEM Nagpur, India as an Assistant Professor. His research area includes parallel computing and algorithms on multicore and manycore architectures.

