Event Based Sentiment Analysis of Twitter Data

Mamta Patil[#1], H.K. Chavan[*2]

[#]*Information Technology, Terna Engineering College, Mumbai University, Nerul, Navi Mumbai, India*

[1]mamtap20@gmail.com

[2]chavan.hari@gmail.com

Abstract- **Everyday large volumes of data are produced. Millions of users share and dissipate most up-to-date information on twitter. Traditional text mining suffers severely from short and noisy nature of tweets. Event detection from twitter data has many new challenges when compared to event detection from traditional media. Noisy nature and limited length are the challenges imposed by twitter data. Event detection performance on twitter is negatively affected by nature of tweets. This paper proposes SegAnalysis framework to tackle these challenges. It performs tweet segmentation, event detection and sentiment analysis. Tweet segmentation is performed in a batch mode using POS (part of speech) tagger on recent online tweets fetched by the user. Segmentation of a tweet preserves the named entities and its stickiness score is calculated. Naïve Bayes classification and online clustering detect events. These events improve situational awareness and decision support. Sentiment analysis categorizes tweets as positive, negative and neutral depending on sentiment score of a tweet.SegAnalysis framework can be extended to deal with events belonging to multiple clusters.**

Keywords: **Tweet Segmentation, Event Detection, Sentiment Analysis, SegAnalysis, Naïve Bayes Classification, Online clustering, Sentiment score.**

I. INTRODUCTION

In the past few years, there has been huge growth in the use of micro blogging platform. Spurred by that growth, companies and media organizations are increasingly seeking ways to mine Twitter. It aims to gain information about what people think and feel about their products and services. Internet users express their opinions, views through tweets. They are attracted to the micro blogging services because of free format of tweets and easy accessibility. Main purpose of tweets is information sharing and communication. These tweets have to be recognized by applications such as Named Entity Recognition (NER), opinion mining and sentiment analysis etc. Twitter length has the limit of 140 characters. It leads to excessive use of abbreviations in tweets which results in noisy tweets. Natural Language Processing (NLP) suffers severely from the noisy and short nature of tweets. Existing classification technique loses the semantic meaning while evaluating such tweets. Hence these tweets need to be segmented intelligently to correctly interpret users' sentiments.

Tweet Segmentation preserves the meaning of tweets. An example of a tweet in Fig.1. "Teacher said to write a letter to your mother on the eve of mother's day." can be split into nine candidate segments among which Teacher said, write a letter, mother and mother's day are preserved as they are semantically meaningful segments. Dividing the tweets into meaningful segments helps in named entity recognition and event detection.

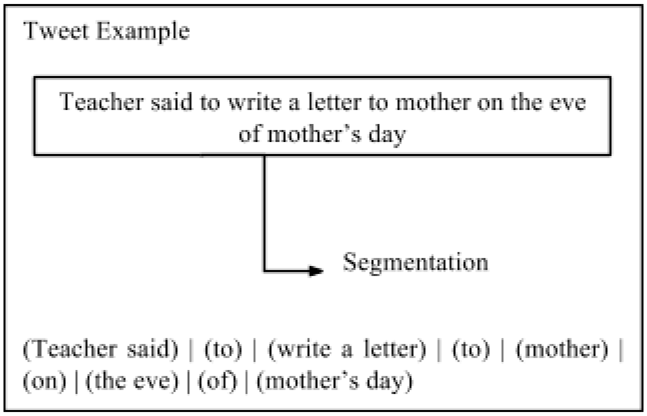

Tweet Example
Teacher said to write a letter to mother on the eve of mother's day
Segmentation
(Teacher said) | (to) | (write a letter) | (to) | (mother) | (on) | (the eve) | (of) | (mother's day)

Fig.1. Example of tweet segmentation

Event detection in twitter refers to identification of events over period of time by analyzing the twitter streams. Event can be defined as "an occurrence causing change in the volume of text data that discusses the associated topic at a specific time" [1]. Twitter is the widely used medium where users share or express their feelings or opinion about real time events such as sports, weather, news, entertainment etc. Twitter attributes such as Topic, Time, People and Location gives details about an event. Detection of real time events help in identifying breaking news or disaster. It also assists in analyzing sentiments of users about specific event.

Sentiment analysis plays an important role in business intelligence. Opinion mining is used by consumers and companies to investigate the sentiment of products before purchase and to monitor the public sentiment of their brands. It is the most dynamic research area in Natural language processing and assists the learning of users' sentiments, views and emotions. Sentiment analysis is the process of automatically extracting knowledge from tweets about any event. The vital task in sentiment analysis of tweets is to categorize them as positive, negative and neutral. Tweets used for sentiment analysis are fetched online in a batch mode. These batches contain tweets related to varied events. Hence, sentiment analysis of such batches has limitations. This paper proposes SegAnalysis framework to improve accuracy by initially detecting events and then evaluating sentiments for specific events.

Remainder of the paper is organized as follows**.** Section II outlines tweet segmentation, event detection and sentiment analysis. The proposed framework is described in section III. Section IV gives the details about optimal tweet segmentation process. Naïve Bayes Classification and online clustering is explained in section V. Section VI briefs about the sentiment analysis. Section VII concludes this paper.

II. RELATED WORK

Natural Language Processing (NLP) is widely used in opinion mining. It consists of tweet segmentation and Named entity recognition (NER). Tweets being short and noisy, several NLP methods used for formal text corpus were not much effective. Many researchers introduced some new techniques along with traditional techniques to improve the performance of tweet segmentation. Using conditional random field CRF model and brown clustering for conventional and tweet specific features Ritter et al. trained a POS tagger [2]. Gimple et al.[3] used new labeling scheme to integrate tweet specific features such as hash tags, URLs and emotions etc. Supervised learning is applied in [4] to identify ill formed words.

T-NER [2] segments named entities with orthographic, contextual, dictionary and tweet-specific features. Named entities are tagged by applying Labeled-LDA with the external knowledge base Freebase. Two stage prediction aggregation model is proposed in [5] for NER. In the first stage, word-level classification is performed by KNN-based classifier. In the second stage, linguistic features are fed into a Conditional Random Field model for fine-grained classification. Community based classification has been used by Chua et al. [6] to extract noun phrases from tweets.

Petrovic et al. [7] proposed Locality Sensitive Hashing to form a story from incoming tweets. This approach does not differentiate whether the new event is news, local event and natural disaster. To detect the disastrous events such as earthquake, Sakaki et al. [8] developed a probabilistic spatiotemporal model. In this model, users have to know the event in advance.

To identify different types of real-world events and their associated social media documents, Becker et al. [9] proposed an online clustering framework. This technique groups together alike tweets and categorize them as event and non-event tweets. Erratic events are excluded by this framework.

Pang and Lee presented overview of existing methods on opinion mining and sentiment analysis [10]. In opinion mining, comparatively less work has been done on micro-blogging. Corpora for sentiment analysis are constructed from web blogs [11] and emoticons assigned to blog posts are used to indicate users' mood.

In [12], emoticons, instead of text, form a training set for the sentiment classification. The dataset was divided into "positive" (texts with happy emoticons) and "negative" (texts with sad or angry emoticons) samples. Using SVM and Naıve Bayes, Emoticonstrained classifiers were able to obtain up to 70% accuracy on the test set. Similar to [12], Go et al obtained 81% accuracy using Naıve Bayes classifier with a mutual information measure for feature selection [13]. However, when three classes are defined ("negative", "positive" and "neutral"), the performance degrades.

To overcome many of the above limitations and challenges, we propose a novel framework SegAnalysis: event based sentiment analysis. By using Naïve Bayes Classification and online clustering, events are detected and sentiment analysis is performed.

III. SEGANALYSIS FRAMEWORK

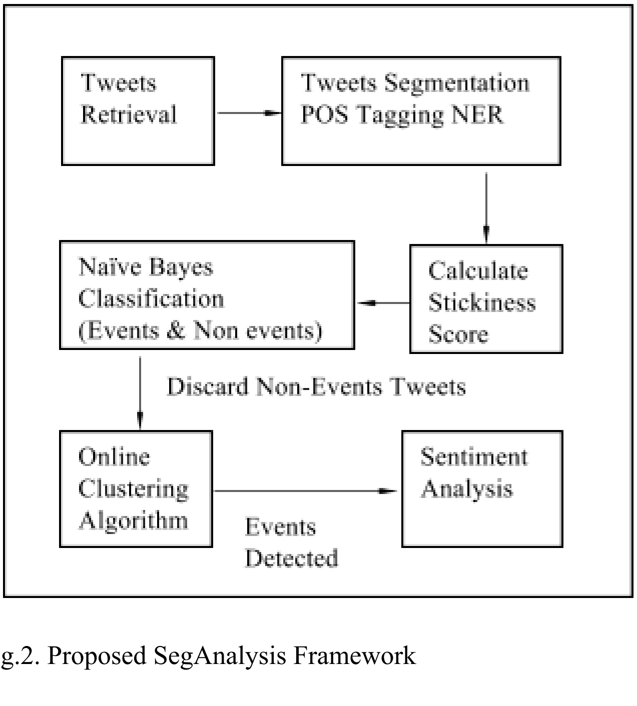

Tweets Retrieval
Tweets Segmentation POS Tagging NER
Calculate Stickiness Score
Naïve Bayes Classification (Events & Non events)
Discard Non-Events Tweets
Online Clustering Algorithm
Events Detected
Sentiment Analysis

Fig.2. Proposed SegAnalysis Framework

SegAnalysis framework is proposed for sentiment analysis using event detection as shown in Fig.2. SegAnalysis consist of tweet segmentation, Naïve Bayes Classification, Online Clustering and Sentiment Analysis. Tweet segmentation extracts meaningful candidate segments from a tweet. It involves removal of stop words, tokenization, spell checking, and POS tagging. POS tagger categorizes candidate segments as noun, adjective, verb, adverb etc. The task of NER is to classify candidate segments into predefined categories such as names of organizations, persons, quantities, monetary values, expressions of times, percentages. Optimal segmentation is achieved by computing stickiness score detailed in Section IV.

Classification integrated with clustering is used to detect the events from twitter streams. Using Naïve Bayes classification, tweets are categorized as non-event and event tweets. Features are computed to extract similar characteristics from tweets. Online clustering algorithm is applied to estimate incoming tweet's similarity to the existing clusters and allocate a tweet to suitable event based cluster. Event detection using Naïve Bayesian Classification and online clustering is discussed elaborately in Section V.

After event detection, sentiments are analyzed for each event. The nature of tweets is analyzed by value of sentiment score. It is computed for a tweet by adding the sentiment scores of individual candidate segments. This results in better accuracy for sentiment analysis. Section VI gives a brief description about Event based Sentiment Analysis.

Proposed framework assists in conserving and defending semantic meaning of tweets. It detects events and decides polarity of tweets despite their noisy nature.

IV. OPTIMAL TWEET SEGMENTATION

Consider a tweet A from batch B, tweet segmentation is to split the words of A into k consecutive segments, A = x_1, x_2, ….., x_k, where each segment x contains one or more words. Stickiness score is calculated independently for each segment. The segment's stickiness score is high if it appears more than once. HybridSeg model is proposed in [14] for Optimal segmentation and is given by

$$argmax \sum_{j=1}^{k} S(x_j) \qquad (1)$$

Stickiness function is calculated based on three factors: length normalization N(x), segment's presence in Wikipedia W(x) and segments phraseness Pr(x).

Length normalization: Tweet segmentation involves mining significant segments from a tweet. Longer segments are topically more important. Normalized length is given by

N(x) = 1 if |x|=1 where x is the segment

$$N(x) = \frac{|x|-1}{|x|} \; if \; |x| > 1 \qquad (2)$$

Presence in Wikipedia: Wikipedia acts as an external dictionary of valid names or phrases. W(x) is the probability that x is an anchor text in Wikipedia, also known as keyphraseness [15], [16]. Let P(x) be the number of Wikipedia entries where x appears in any form and Pa(x) be the number of Wikipedia entries where x appears in the form of anchor text. W(x) is given by

$$W(x) = \frac{Pa(x)}{P(x)} \qquad (3)$$

Segment phraseness: is probability of x being phrase based on local and global context. It is calculated by Symmetric conditional probability (SCP)

$$SCP(x) = log \frac{Pr(x)^2}{\frac{1}{|x|-1} \sum_{j=1}^{|x|-1} Pr\,(w_1.....w_j)\,Pr(w_{j+1}\,......w_{|x|})} \qquad (4)$$

In above equation, Pr(x) or Pr(w_1….w_j) is the approximated n-gram probability of a segment. If x contains a single word w, then SCP(x) is given by

$$SCP(x) = 2logPr(w) \qquad (5)$$

Stickiness of x, S(x), considers all the three factors explained above. Stickiness of x is given by

$$S(x) = N(x).e^{W(x)} \frac{2}{1+e^{-SCP(x)}} \qquad (6)$$

Stickiness score assist in achieving optimal tweet segmentation.

V. EVENT DETECTION USING NAÏVE BAYES CLASSIFICATION AND ONLINE CLUSTERING

Naïve Bayes classification model is simple yet very powerful. It is used in this framework to separate event related tweets from nonevent content. A tweet t is represented by a vector s_1, s_2…s_k of binary or weighted segments. The Probability that t is an event [1] is denoted by P(E|s_1,s_2…..s_k) can be written as follows using Bayes theorem:

$$P(s_1, s_2, ……s_k) = P(E).(\frac{P(s_1,s_2,…s_k|E)}{P(s_1,s_2,…s_k)}) \qquad (7)$$

Similarly, the Probability that t is a non- event can be written as:

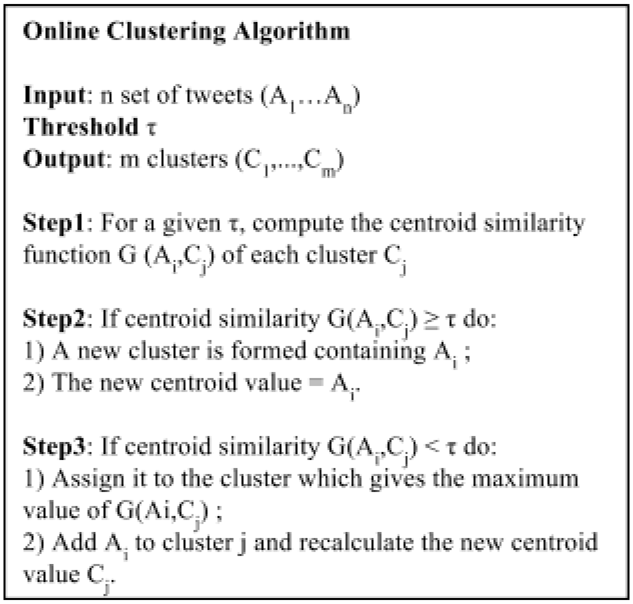
$$P(s_1, s_2, ……s_k) = P(N).(\frac{P(s_1,s_2,…s_k|N)}{P(s_1,s_2,…s_k)}) \qquad (8)$$

Using the assumption of independence among the segments in t as well as our prior calculations of P(E), P(N), P(s_i|E), and P(s_i|N), we introduce the threshold (D) :

$$D{=}log \frac{P(N|s_1,s_2,…s_k)}{P(E|s_1,s_2,…s_k)} = log \frac{P(N)}{P(E)} + \sum_{i}^{k} log \frac{P(s_i|N)}{P(s_i|E)} \qquad (9)$$

Tweets with a positive value of D are classified as non-events and discarded. Others are treated as events.

Online clustering algorithm is applied to event related tweets to identify the event. In online clustering, the numbers of clusters are not pre-defined nor are any detail related to features of real-world events available. Yet this algorithm automatically assigns each tweet to a cluster. An event is a vector in which each dimension is the probability of some feature in the event. Textual content of tweet is used to represent tweet as a TF-IDF weight vector. A cosine similarity metric is applied as the centroid similarity function G.



```
Online Clustering Algorithm

Input: n set of tweets $(A_1 \ldots A_n)$
Threshold $\tau$
Output: m clusters $(C_1,\ldots,C_m)$

Step1: For a given $\tau$, compute the centroid similarity
function $G(A_i,C_j)$ of each cluster $C_j$

Step2: If centroid similarity $G(A_i,C_j) \geq \tau$ do:
    1) A new cluster is formed containing $A_i$ ;
    2) The new centroid value = $A_i$.

Step3: If centroid similarity $G(A_i,C_j) < \tau$ do:
    1) Assign it to the cluster which gives the maximum
       value of $G(Ai,C_j)$ ;
    2) Add $A_i$ to cluster j and recalculate the new centroid
       value $C_j$.
```

Fig.3. Online Clustering Algorithm [1]

A set of features $(F_1, ..., F_k)$ for each tweet $(A_1, ..., A_n)$ is used to compute the cosine similarity measure between the tweet and each cluster $(C_1, ..., C_m)$. The similarity function is computed against each cluster C_j for $j = 1, . . . , m$ where m is the number of clusters (initially $m = 0$).

Centroid similarity function $G(A_i, C_j)$ of a cluster is calculated through average weight of each tweet segment in respective cluster.

During the training phase, the threshold parameters (τ) are determined pragmatically. Based on the tweet's similarity to the existing cluster, online clustering algorithm determines the suitable cluster assignment for each tweet.

A new cluster C_m is created when there is no cluster whose centroid similarity function $G(A_i,C_j)$ is greater than τ.

Each cluster created by online clustering algorithm represents an event.

VI. EVENT BASED SENTIMENT ANALYSIS

Sentiment analysis evaluates the polarity of tweets i.e. positive, negative and neutral. For example, a strong positive opinion can be inferred from the word "excited" whereas the word "depressed" possesses a strong negative association. In SegAnalysis, sentiments are analyzed for a specific event. A publically available lexical resource SentiWordNet is used to calculate Sentiment Score for each candidate segment of a tweet. Summation of individual Sentiment Score of candidate segments gives the Sentiment Score for a tweet. A positive value, negative value and zero value of Sentiment Score for a tweet indicates positive polarity, negative polarity and neutral polarity, respectively. Using Tweet Sentiment Score, sentiment for specific event can be inferred. Sentiment analysis for particular events gives better accuracy.

VII. CONCLUSION

Event detection plays an important role in sentiment analysis of twitter data. Segmentation preserves semantic meaning of tweets which benefits NER (Named Entity Recognition). SegAnalysis fetches online tweets and segments them into meaningful candidates. Combined mechanism of Naïve Bayes and online clustering abets identification of unknown events over period of time. Naïve Bayes reduces computational overhead by extracting only event related tweets. Online clustering detects events in real time. Nature of detected events is evaluated using sentiment score. SegAnalysis framework can be extended to deal with events belonging to multiple clusters.

REFERENCES

[1] Alsaedi, N., Burnap, P. and Rana, "A Combined Classification-Clustering Framework for Identifying Disruptive Events", Proceedings of 7th ASE International Conference on Social Computing, pp. 1–10, 2014.

[2] A. Ritter, S. Clark, Mausam and O. Etzioni, "Named entity recognition in tweets: An experimental study", in Proceedings Conference on Empirical Methods Natural Language Process, pp. 1524–1534, 2011.

[3] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments", in Proceedings 49th Annual Meeting the Association for Computational Linguistics (ACL):Human Language Technology, pp. 42–47, 2011.

[4] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a #twitter", in Proceedings. 49th Annual Meeting Association for Computational Linguistics: Human Language Technol., pp. 368–378, 2011.

[5] X. Liu, S. Zhang, F. Wei and M. Zhou, "Recognizing named entities in tweets", in Proceedings. 49th Annual Meeting Association for Computational Linguistics: Human Language Technol., pp. 359–367, 2011.

[6] F.C.T. Chua, W. W. Cohen, J. Betteridge and E. -P. Lim, "Community-based classification of noun phrases in twitter", in Proceedings 21st ACM

International Conference on Information and Knowledge Management, pp. 1702–1706, 2012.

[7] Petrović, S., Osborne, M. and Lavrenko, “Streaming first story detection with application to twitter”, Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 181–189, 2010.

[8] Sakaki, T., Okazaki, M. and Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors”, 19th International World Wide Web Conference (WWW ’10), 2010.

[9] Becker, H., Naaman, M. and Gravano, “Beyond Trending Topics: Real-World Event Identification on Twitter”, ICWSM, pp. 1–17, 2011.

[10] B. Pang and L. Lee, “Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval”, 2(1-2):1{135, 2008.

[11] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen, “Emotion classification using web blog corpora”, In WI ’07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA. IEEE Computer Society, pages 275–278, 2007.

[12] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification”, In Proceedings of the ACL Student Research Workshop, pages 43, 48, 2005.

[13] A. Go, L. Huang, and R. Bhayani, “Twitter sentiment classification using distant supervision”, In CS224N Project Report, Stanford, 2009.

[14] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, “Tweet Segmentation and Its application to Named Entity Recognition”, in Proceedings IEEE Transactions on knowledge and data enginerring vol. 27, No. 2, February 2015

[15] D. N. Milne and I. H. Witten, “Learning to link with wikipedia”, in Proceedings 17th ACM International Conference on Information and Knowledge Management, 2008, pp. 509–518.

[16] R. Mihalcea and A. Csomai, “Wikify!: linking documents to encyclopedic knowledge”, in Proceedings. 16th ACM Conference on Information and Knowledge Management, 2007, pp. 233–242.