RESEARCH ARTICLE OPEN ACCESS

Efficient Density Based Clustering of Tweets and Sentimental Analysis Based on Segmentation

AnumolBabu[1], Rose V Pattani[2]
[1](Post Graduate Student, Dept. of CSE, Mangalam College of Engineering,Kerala, India)
[2] (Assistant Professor, Dept. of CSE, Mangalam College of Engineering,,Kerala, India)

----------**********----------

Abstract:

Twitter has become popular social networking site where users share their up-to-date information. The error-prone and short nature of tweets makes the word-based representation less reliable. Tweet segmentation is the process of splitting tweets into meaning segments so that its semantic meaning is well conserved and is easy to be used by downstream applications. Segmentation is done based on stickiness score considering both global and local context. Clustering of tweets are done using DBSCAN method with Jaccard Coefficient as the similarity measure. The sentimental variations in tweets are measured based on segmentation. The experimental evaluation shows that the global terms using wikilinks are more efficient than the normal segmentation. Clustering is more effective using DBSCAN algorithm, which is best for uncertain data.

Keywords — **Segmentation, stickiness score, named entity recognition, clustering, sentiment analysis.**

----------**********----------

I. INTRODUCTION

In the past few years there has been an exponential rise in the use of online social media systems like Twitter. It has become one of the most important platform for people to find, share, and publish timely information. Tweets are short messages, limited to 140 characters in length. Due to its large volume of timely information generated by its millions of users, it is important to understand tweets' language for a large body of downstream applications, such as named entity recognition [1],[2],[3], sentimental analysis, opinion mining etc.

Due to the length limitation and no constraints on its writing styles often word abbreviations are used, and in other cases words are misspelled or contain grammatical errors. The error-prone and short nature of tweets often make the word-level representation model less reliable. A segment-based tweet representation model has been proposed .A segment can be a named entity, a semantically meaningful information unit, or any other types of phrases which appear "more than by chance".

Data redundancy is one of the important problem of Twitter. In twitter users are likely to generate similar tweets about some popular topics/events. By clustering these similar tweets together, we can generate a more short and structured representation of the collection of tweets, which will be very useful for many Twitter-based applications (e.g., trend analysis). Clustering is a standard data mining task which requires two important components: a distance metric to find the similarity between data points and a clustering algorithm that merges data points into different clusters based on the similarity characterized by the distance metric.

Sentiment analysis of dataset is considered as a much harder problem than that of conventional text

such as review documents. This is mainly due to the short length of tweets, the frequent use of informal and irregular words, and the rapid evolution of language in Twitter.

II. RELATED WORK

Tweets are well-known for their error-prone and short nature. This leads to failure of many conventional NLP techniques, which depend on local linguistic features, such as capitalization of words, POS tags of previous words, etc

For named entity recognition in tweets both supervised and unsupervised methods have been proposed. HybridSeg framework jointly utilizes both the local context knowledge and the global knowledge bases in the process of tweet segmentation which benefits many downstream applications. It finds the optimal segmentation by maximizing the stickiness score [1],[3]. T-NER is a part of the tweet-specific NLP framework [4]. T-NER first segments named entities using a CRF model with orthographic, contextual, and dictionary features, and then labels the named entities using a LDA (Latent Dirichlet allocation) model. The NER solution proposed in [5] is also based on a CRF model. combine a K-Nearest Neighbors (KNN) classifier under a semi-supervised learning framework. The unsupervised approach named TwiNER recognizes named entities with two steps: tweet segmentation and ranking of segment. In TwiNER global context obtained from Wikipedia and Web N-Gram corpus jointly used to partition tweets into valid segments (phrases) [2]. Tweet segmentation is conceptually similar to Chinese word segmentation (CSW) [9]. Text in Chinese is a continuous sequence of characters. Segmenting the sequence into meaningful words is the first step in most applications. CSW methods are mostly developed using supervised learning techniques like perceptron learning and CRF model.

Clustering uncertain data has been well known as an important issue. The density-based clustering methods like DBSCAN is used to cluster uncertain data, by considering the geometric distances between objects [6]. A Jaccard index based clustering algorithm (JIBCA) support mining online reviews and predicting sales performance [7]. It is a clustering and regression based algorithm for online data sentiment prediction. The two Latent Dirichlet Allocation (LDA) based models: Foreground and Background LDA (FB-LDA) and Reason Candidate and Background LDA (RCB-LDA) [8] assess public sentiment variations on Twitter and extract possible reasons behind such variations. Emoticon Smoothed Language Model (ESLAM) [10] is a probabilistic language model used for sentiment analysis in twitter that train based on the manually labelled data, and then use the noisy emoticon data for smoothing,

III. PROPOSED SYSTEM

The proposed system segments tweets in batch mode. Tweets from a targeted Twitter stream are first collected and they are grouped into batches by their publication time using a fixed time interval (e.g., a day). Each batch of tweets are then segmented collectively.

The architecture of the proposed system is shown in Fig 1. For tweet segmentation tweet dataset is preprocessed by removing stopwords applying stemming and preprocessing. Tweet segmentation is done based on the stickiness score calculation. Stickiness score depends on the three factors - Length Normalization, Key Phraseness and Segment Phraseness. Named Entity Recognition by Random Walk and POS Tag method is done using segments. Clustering is done by DBSCAN algorithm using Jaccard Similarity measure. Also sentimental variations in tweets are analyzed based on segmentation

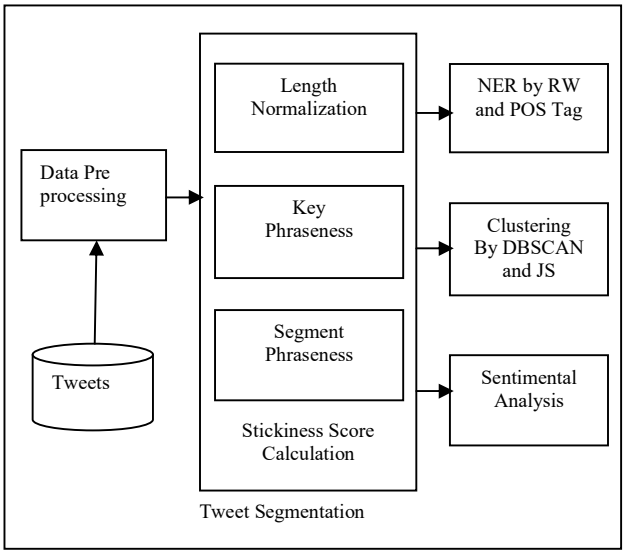

Fig 1. Architecture

A. *Tweet Segmentation*

Given a tweet t from batch T, the idea of tweet segmentation is to split the l words in t into m < l' consecutive segments, where each segment contains one or more words. The optimal segmentation of a tweet is calculated by using the stickiness score value. A high stickiness score of segment s indicates that it is a phrase which appears "more than by chance", and more splitting of the segment could break the semantic meaning of the phrase.

The stickiness score of a segment is taken by three factors:
(i) Length normalization L(s)
(ii) Segment's presence in Wikipedia Q(s)
(iii) Segment's phraseness or the probability of s being a phrase based on global and local contexts Pr(s)

1) ***Length Normalization.:*** The basic idea of tweet segmentation is to extract meaningful phrases. The longer segments are commonly chosen which contain more helpful information. Let |s| be the number of words in segment s, then the normalized segment length is calculated as L(s) = 1 if |s|= 1 and L(s) =(|s|-1)/|s| if |s|> 1.

2) ***Presence in Wikipedia:*** Wikipedia serves as an external dictionary of valid names or phrases for segmentation. The probability that a segment is an anchor text in Wikipedia is also known as key phraseness. Let wiki(s) and $wiki_a(s)$ be the number of Wikipedia entries where the segment appears in any form and the segment appear in the form of an anchor text, respectively, Q(s) = $wiki(s)/wiki_a(s)$. The segment that is frequently used as an anchor text in Wikipedia is chosen for segmentation

3) ***Segment phraseness:*** The Segment's phraseness is estimated as the probability of a segment being a valid phrase using Symmetric Conditional Probability (SCP) measure. It takes n-gram probability of a segment for calculation

B. *Named Entity Recognition*

The two segment-based NER methods are used. The first one identifies named entities from a collection of segments by considering the co-occurrences of named entities. The second one does so based on the POS tags of the constituent words of the segments.

1) ***NER by Random Walk***: This method works based on the observation that a named entity frequently co-occurs with other named entities in a batch of tweets (i.e., the gregarious property). Based on this observation, a segment graph is constructed first. A node in this graph is a segment identified by segmentation. An edge exists between two nodes if they co-occur in some tweets. A random walk model is then applied to the segment graph. The top K segments are chosen as named entities.

2) ***NER by POS Tagger***: This method uses the part of speech tags in tweets and noun phrases are considered as named entities using segment instead of word as a unit. A segment may appear in different tweets and words contained in it may be assigned different POS tags in these tweets. The

evaluation of the likelihood of a segment being a noun phrase by considering the POS tags of its constituent words of all appearances are done
.

C. Clustering

Clustering is an unsupervised learning technique in which a collection of objects such as tweets are taken and they are organized into groups based on their similarity. The groups that are formed are known as clusters. In density-based clustering methods clusters are considered as dense regions of objects that are separated by regions of low density. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is one of the most common density-based clustering. The algorithm mainly requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts). The set of points taken for clustering is divided into - core points, border points and outliers. The algorithm starts with an arbitrary starting point. If it is a core point, then it forms a cluster together with all points (core or non-core) that are reachable from it otherwise it is labeled as noise. The algorithm iteratively examines every object in the dataset until no new object can be added to any cluster.

Jaccard Coefficient is a statistical measure for finding the similarity between documents, or binary data. It is defined as the size of intersection between the datasets divided by the size of the union of the datasets.

D. Sentimental Analysis

Sentimental analysis is the analysis of the feelings (i.e. attitudes, emotions and opinions) behind the words. In sentiment analysis the opinions are classified into positive, negative, or neutral. The Sentistrength Tool is used for finding the positive or negative score for tweets .The tool is based on the **L**inguistic **I**nquiry and **W**ord **C**ount (LIWC) sentiment lexicon. Basically, it reads a given text and counts the percentage of words that reflect different emotions, thinking styles, social concerns, and even parts of speech. Sentistrength tool works in the following way: At first each word in the text has assign a sentiment score according to the sentiment lexicon. Then the maximum positive score and the maximum negative score is selected among all individual words in the text, and the sum of the maximum positive score and the maximum negative score is denoted as FinalScore. Finally the sign of FinalScore is used to indicate whether a tweet is positive, neutral or negative.

IV. EVALUATION

We have setup an experiment to perform the tweet segmentation and its accuracy is evaluated. The quality of segmentation in local context as well as global contact is learned. From the evaluation between learning from weak NERs and learning from local collocation, it is found that the global terms used as anchor text in wikipedia are more efficient than the ordinary segmentation. Also the clustering, as an enhancement to the present system is found to be effective, when a DBSCAN algorithm is used.

For our experiment, we have used synthetic dataset generated through a simulated Twitter application. The tweets collected from users are preprocessed to make the input of the experiment. The StandFord NLP Library is used for applying Natural language Processing techniques to find semantics. The Sentistrength Tool is used for finding the positive or negative score for tweets. The overall stickiness score is calculated and the tweets beyond a threshold Lambda=0.9 is taken for segmentation. The ordinary clustering method was replaced with a DBSCAN algorithm, which is well suited for uncertain data.

Fig.2 represents the comparison of no of segments vs no of clusters in different dataset. It can be shown that as the no. of tweets are increased, the no of clusters are also increased. Fig 3 represents the precision of segmentation vs clustering and Fig 4 represents the recall of segmentation vs clustering.

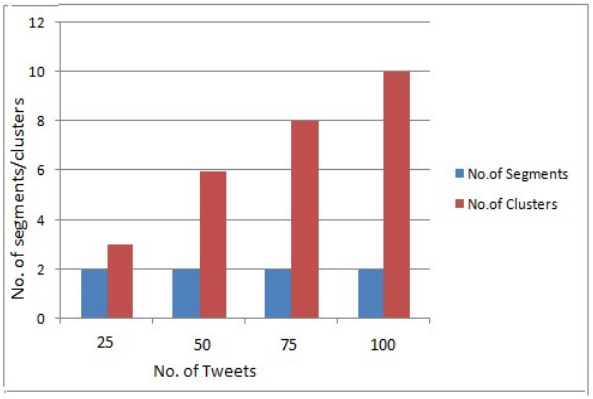

Fig 2. No of segments vs No of Clusters

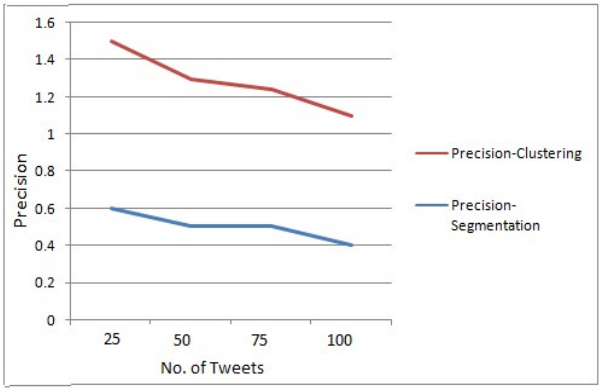

Fig 3. Precision of Segmentation vs Clustering

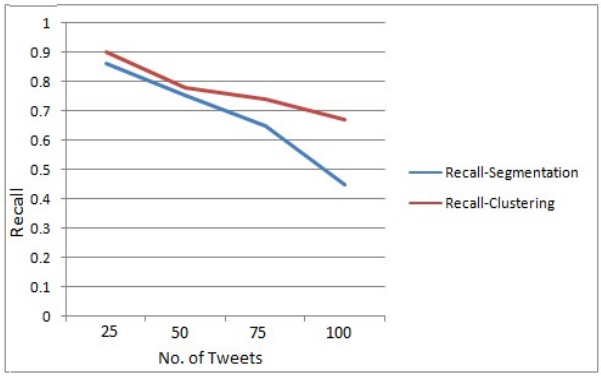

Fig 4. Recall of Segmentation vs Clustering

V. CONCLUSIONS

Twitter has attracted millions of users to share their interests, opinions etc. The limited length of tweets makes it difficult to understand their meaning. Tweet segmentation splits the tweet into a number of meaningful segments. The local context and global context along with their stickiness value is considered in segmentation process. DBSCAN algorithm with Jaccard Similarity measure is used for clustering of tweets. Clusters are very helpful in many twitter based applications. Density based clustering methods are best for uncertain data and clustering is efficient. By applying sentimental variations tweets can be effectively classified into positive or negative. In future the segment-based representation can be used for other tasks like event detection, opinion mining etc and we can consider multiple linguistic factors in segmentation process.

REFERENCES

[1] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi H, "Tweet Segmentation and Its Application to Named Entity Recognition" *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 2, FEBRUARY 2015*

[2] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 721–730.

[3] C. Li, A. Sun, J. Weng, and Q. He, "Exploiting hybrid contexts for tweet segmentation," in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 523–532

[4] .X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol., 2011, pp. 359–367.

[5] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In Proc. of EMNLP, 2011.

[6] Bin Jiang, Jian Pei, Yufei Tao, "Clustering Uncertain Data Based on Probability Distribution Similarity," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013

[7] Nihalahmad R. Shikalgar, Arati M. Dixit, "JIBCA: Jaccard Index based Clustering Algorithm for Mining Online Review", *International Journal of Computer Applications (0975 – 8887) Volume 105 – No. 15, November 2014*.

[8] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu,, Chun Chen "Interpreting the Public Sentiment Variations on Twitter," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 5, MAY 2014.

[9] Y. Zhang and S. Clark, "A fast decoder for joint word segmen- tation and pos-tagging using a single discriminative model," in Proc. Conf. Empirical Methods Natural Language Process., 2010, pp. 843–852.

[10] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1678–1684.