# Online Tweet Summarization and Ranking for Named Institution Impression

**Rani Dubey[1], Amitav Saran[1]**

*[1]Assistant Professor, [1]Dept. of CSE*

*[1]Gandhi Institute for Technology, Bhubaneshwar, India*

## Abstract

Number of private and public Institution is reported to create and impression targeted Twitter streams to collect and understand user's opinions about the organizations. Big data is analyzed with the software methods commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. An effective tweet summarization clustering algorithm is access for effective clustering of tweets with only one time pass over the data. This algorithm uses two data analysis model is important tweet information in clusters. Clustering of tweets is done using DBSCAN method with Jacquards Coefficient as the similarity methods. The tweets summarization is increased based on segmentation. The experimental evaluation shows that the global terms using wiki links are more efficient than the normal segmentation. Clustering is very effective in DBSCAN algorithm is find uncertain data. Further tweets are strongly effective with their posted new messages tweets is very fast rate. Classification models establish the optimal cluster of a tweet is task in terms of tweet cluster vector. We can implement in real time tweet environments to identify the rumor with high level security.

**Index Terms:** Segmentation, named entity recognition, clustering, sentiment analysis, Tweet Summarization, Linguistic features, Feature Extraction.

## INTRODUCTION

Micro blogging services such as Twitter has extracted millions of users to share their information between the people and extract knowledge from the shared information, [2] as they offer large volumes of real time data, with around 200 millions of regular users posting the tweets per day in June 2015. The people upload person information but they may get used with that information. For this, the users must be able to understand the tweets so that they can gain some knowledge and also continue their comments on the same topic. 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 [3],[4],[5], 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. We can generate a more short and structured representation of the collection of tweets, which will be very useful for many Twitter-based applications [6]. 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.

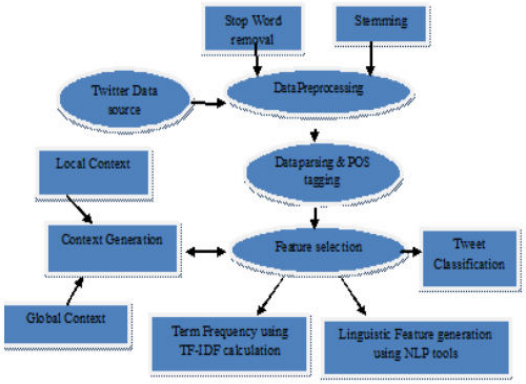


Fig.1 System for Tweet Segmentation

Data Classification Technique for Text Message Many short message classification techniques rely on the local and global context with respect to the linguistic features of the message which are as follows [7].

- K- Nearest Neighbors: KNN-based classifier is used to conduct word-level classification, leveraging the similar and recently labeled tweets.
- Conditional Random Field (CRF): It is class of statistical modeling method often applied in pattern recognition and machine learning, where they are used for structured prediction. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data, such as natural language text or biological sequences and in computer vision.

## 2. Related Work

Tweets are infamous for their error-prone and short nature. This leads to failure of many conventional NLP techniques. Also acknowledging the error-prone nature of tweets, Han and Baldwin [8] proposed to normalize ill formed words in tweets to make the contents more formal. The speed improvement is achieved by the use of a single-beam decoder. Given an input sentence, candidate outputs are built incrementally, one character at a time. When each character is processed, The idea is to use the POS of a partial word as the predicted POS of the full word it will become. Possible predictions are made with the first character of the word [9]. Clustering uncertain data has been well known as an important issue. The density-based clustering methods like DBSCAN are used to cluster uncertain data, by considering the geometric distances between objects [10]. A Jaccard index-based clustering algorithm (JIBCA) support mining online reviews and predicting sales performance [11]. 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 (RCBLDA) [12] assess public sentiment variations on Twitter and extract possible reasons behind such variations. Emoticon Smoothed Language Model (ESLAM) is a probabilistic language model used for sentiment analysis in twitter that train based on the manually labeled data, and then use the noisy emoticon data for smoothing. Silviu Cucerzan [13] proposed large-scale system for the named entity recognition and semantic disambiguation based on information extracted from a large encyclopedic collection and Web search results.. This system treat mention detection and entity disambiguation as two different problems. Milne and Witten [7] proposed system that describes how to automatically cross-reference documents with Wikipedia. It explains how machine learning can be used to identify significant terms within unstructured text, and enrich it with links to the appropriate Wikipedia articles.

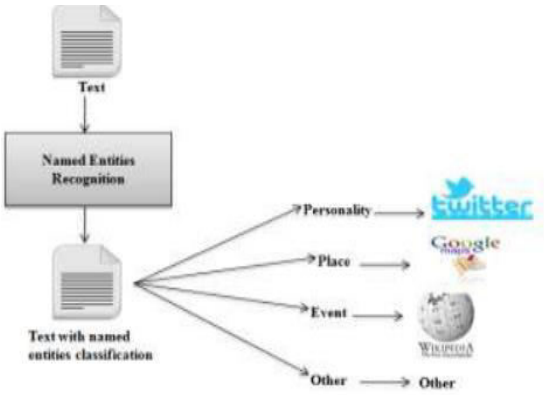


Fig. 2 System with entity classification & linking

## 3. System Architecture

Twitter is a micro-blogging platform has become a major social media platform with hundreds of millions of users. Twitter is a social network where users can publish and exchange short messages of up to 140 characters long, also known as tweets. [14].We define a rumor to an unverified assertion that starts from one or more sources and spreads over time from node to node in a network. There are usually several rumors about the same topic, any number of which can be true or false. Twitter datasets are collected and stored datasets as collected in big database. The data discovery platform is used to extract the key features from uploaded datasets. The keywords analyzed based POS tagger.[15] after that analysis portfolio is used to predict the sentiments and labeled as positive and negative. It can be stored enterprises data warehouses. Business portfolio is used to predict the rumors based on KNN classifiers. KNN classification approach is used to label the each tweets.

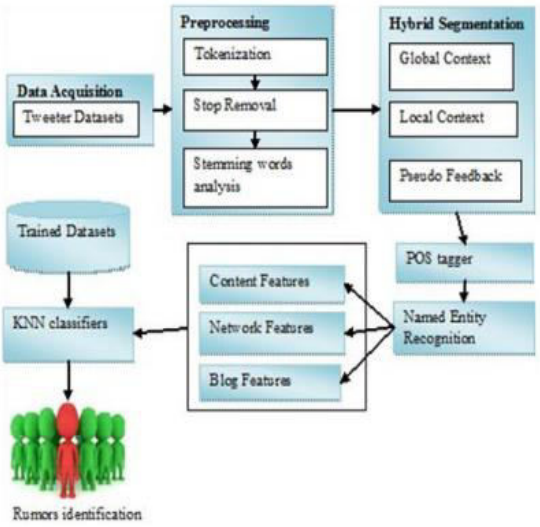


Fig.3 Overall System Architecture

## 4. Proposed System

The proposed system segments tweets in batch approach from a targeted Twitter stream is first collected and they are grouped into batches in their publication time using a fixed time interval. Each batch of tweets is segmented collectively, and architecture of the proposed system is tweet segmentation tweet dataset is preprocessed by removing stop words applying stemming and preprocessing. Tweet segmentation is based on the stickiness score calculation. Stickiness score depends on the three factors. Length Normalization, Key Phrasings and Segment Phrasings. Named Entity Recognition by Random Walk and POS Tag method is done using segments. Clustering is done by DBSCAN algorithm using Jacquards Similarity measure. Also sentimental variations in tweets are analyzed based on segmentation

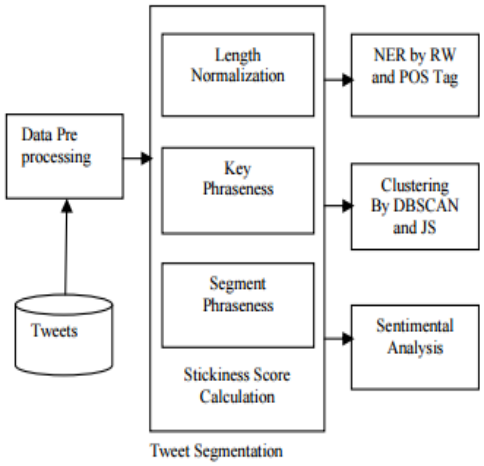


Fig. 4 Architecture

### A. Tweet Segmentation

Given 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 many words. The optimal segmentation of a tweet is calculated by using the stickiness score value. 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 phrasings or the probability of s being a phrase based on global and local contexts Pr(s)

### B. Named Entity Recognition

The two-segment based NER model is used. The first one identifies named entities from a collection of segments by considering the concurrences of named entities. The second one is based on the POS tags of the constituent words of the segments.

### C. Clustering

Clustering is an unsupervised learning approaches in collection of objects such as tweets is taken and organized into groups-based points taken for clustering is divided into core points border points and outliers. The algorithm starts with an arbitrary starting point it is a core point then it forms a cluster together with all points that are reachable from it otherwise it is labeled. The algorithm iteratively examines every object in the dataset until old object can be added to any cluster. Jaccard Coefficient is a statistical measure for dictions of the similarity documents or binary data. It is defined as the size of intersection among the datasets divided by the size of the same datasets.

### D. Sentimental Analysis

Sentimental analysis is feelings behind the words in sentiment analysis the opinions is classified into positive, negative, or neutral. The Sent strength Tool is used for finding the positive or negative score tweets .The tool is based on the Linguistic Inquiry and Word Count (LIWC) sentiment lexicon. Basically, the maximum positive score and the maximum negative score is selected many individual words in the text and the sum of the maximum positive score and the maximum negative score is denoted as Final Score. Finally, the sign of Final Score is used to indicate whether a tweet is positive, neutral or negative.

## 5. Rumor Identification

The nearer neighbors contribute more to the average distant ones a common weighting method is consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors is taken from a set of objects for which the class the object property value is known. The training vectors in a multidimensional feature space each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class

labels of the training samples. In the classification phase, k is a user-defined constant and an unlabeled vector is classified by assigning the label which is most frequent among the k training samples nearest to that query point. Often the classification accuracy of k-NN can be improved significantly if the distance metric is learned with specialized algorithms such as Neighbor or Neighborhood components analysis.

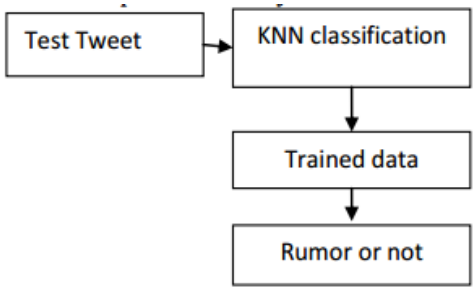


Fig. 5 Rumor identification

**A. Mathematical Model**

KNN algorithm as derived as follows:

BEGIN

Input: D={(X1,C1),….,(Xn,Cn)}

X=(X1,…, Xn) new instance to be classified

For each labeled instance (Xi,Ci) calculate d(Xi,X)

Order d(Xi,X) from lowest to highest, (i-1,…,N)

Select the K nearest instances to X:DX K

Assign to X the most frequent class in DX K

End:

Classify the tweet into different classes based on term frequency and Linguistic features of the tweet data. Many named entities and common phrases are preserved in tweets for information sharing and dissemination. There exist tweets composed in proper English the normalization of the data is carried out using different methods.

## 6. Evaluation

We have setup an experiment to perform the tweet segmentation and its speed and 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 the global terms used as anchor text in Wikipedia is more efficient than the ordinary segmentation. The Stand Ford NLP Library is used for applying Natural language processing techniques to find semantics. The Seniti strength 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.

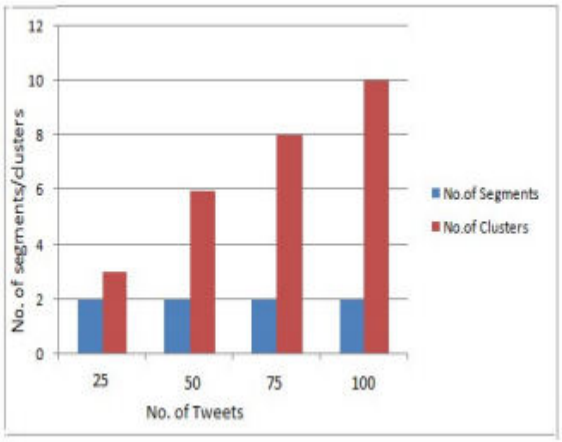


Fig. 6 Number of segments versus Number of Clusters

## 7. Conclusions and Future Work

We designed many features to use in the classification of tweets in order to develop a system through which informational data may be filtered from the conversations is much value in the context of searching for immediate information for relief efforts to utilize in order to minimize damages. The local context and global context many stickiness value is considered in segmentation process. DBSCAN algorithm with Jaccard Similarity measure is used for clustering of tweets. Density based clustering methods is best for uncertain data and clustering is efficient by applying sentimental number of tweets is effectively classified into positive or negative. In future the segment-based representation is used for other tasks like event detection opinion mining and we can consider multiple linguistic factors in segmentation process. Many private and/or public instructions have been reported to change twitter stream to collect and understand users' opinions about the organizations. Nevertheless, it is practically infeasible and unnecessary to listen and modify the total twitter stream in extremely large volume. Future work is extending our method is implement number of classification algorithm to predict the attackers and also eliminate the attackers from twitter datasets and try this model to implement in number of languages in twitter.

## References

[1] 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.

[2] B. Han and T. Baldwin, "Lexical normalisation of short text messages: Makn sens a #twitter," in Proc.49th Annu. Meeting. Assoc.Comput. Linguistics: Human Language Technol., 2011, pp. 368–378.

[3] 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

[4] 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.

[5] 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

[6] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entity centric topic-oriented opinion summarization in twitter," in Proc.18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 379–387.

[7] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in Proc. IEEE 7th Int. Conf. Data Mining, 2007, pp. 697–702.

[8] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In Proc. of ACL, 2011

[9]. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.

[10] 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

[11] 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.

[12] 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.

[13] S. Guo, M.-W. Chang, and E. Kiciman, ―To link or not to link? A study on end-to-end tweet entity linking,‖ in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol., pp. 1020–1030, 2013

[14]. L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In Proc. of CoNLL, 2009.

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