International Journal of Current Research and Modern Education (IJCRME)
ISSN (Online): 2455 - 5428
(www.rdmodernresearch.com) Volume I, Issue I, 2016

TWEET SEGMENTATION AND CLASSIFICATION FOR RUMOR IDENTIFICATION USING KNN APPROACH

R. Gomathi* & M. Rajakumar**

* PG Scholar, Department of Master of Computer Applications, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamilnadu

** Associate Professor & Head, Department of Master of Computer Applications, Dhanalakshmi Srinivasan Engineering College, Perambalur, Tamilnadu

Abstract:

Experimentally, magnetic tape item has been Abstract: Big data analytics is the process of examining large data sets containing a variety of data types i.e., big data to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. In this project, we analyze social media data. Social media analytics is the practice of gathering data from blogs and social media websites and analyzing that data to make business decisions. The most common use of social media analytics is to mine customer sentiment in order to support marketing and customer service activities. And then we take twitter big data to predict named entity. Considering wide use of Twitter as the source of information, reaching an interesting tweet for a user among a bunch of tweets is challenging. In this work, it is aimed to reduce the Twitter user's effort to access to the tweet carrying the information of interest. To this aim, a tweet recommendation method under a user interest model generated via named entities is presented. To achieve our goal, Hybrid Seg is generated via named entities extracted from user's followees and user's own posts. And extend our approach to analyze short text in tweets and rumor based tweets. So we implement KNN approach to eliminate rumor based tweets with improved accuracy rates. We can implement in real time tweet environments to identify the rumor with high level security.

Key Words: Twitter, Tweets, Sentiment Analysis & Named Entity Recognition

1. Introduction:

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. The ubiquity, accessibility, speed and ease-of-use of Twitter have made it an invaluable communication tool. People turn to Twitter for a variety of purposes, from everyday chatter to reading about breaking news. Social networks are popular media for sharing information. Online social networks enable large-scale information dissemination in a very short time, often not matched by traditional media. Mis-information and false claims can also propagate rapidly through social networks. This is exacerbated by the fact that (i) anyone can publish (incorrect) information and (ii) it is hard to tell who the original source of the information is Twitter is a massive social networking site tuned towards fast communication. More than 140 million active users publish over 400 million "Tweets" every day. Twitter's speed and ease of publication have made it an important communication medium for people from all walks of life. Twitter has played a prominent role in socio-political events, such as the Arab Spring and the Occupy Wall Street movement. Twitter has also been used to post damage reports and disaster preparedness information during large natural disasters, such as the Hurricane Sand. While the spread of inaccurate or questionable information has always been a concern,

the emergence of the Internet and social media has exacerbated the problem by facilitating the spread of such information to large communities of users. This is especially the case in emergency situations, where the spread of a false rumour can have dangerous consequences. For instance, in a situation where a hurricane is hitting a region or a terrorist attack occurs in a city, access to accurate information is crucial for finding out how to stay safe and for maximizing citizens' well-being. We introduce a novel methodology to create a dataset of rumors and non-rumors posted in social media as an event unfolds. This methodology consists of three main steps: (i) collection of (source) tweets posted during an emergency situation, sampling in such a way that it is manageable for human assessment, while generating a good number of rumorous tweets from multiple stories, (ii) collection of conversations associated with each of the source tweets, which includes a set of replies discussing the source tweet, and (iii) collection of human annotations on the tweets sampled. To facilitate the annotation task, we developed a tool that visualizes the timeline of tweets associated with an event. The purpose of the tool is to enable annotators to read through the tweets and annotate them as being rumors or non-rumors. Annotators record their selections by clicking on the appropriate icon next to each source tweet (green tick for a rumour, a red cross for a non-rumour, or an orange question mark). Each source tweet is also accompanied by a bubble icon, which the annotator can click on to visualize the conversation sparked by a source tweet.

Re-Tweet: A re-tweet is a repost or forward of a tweet by another user. It is indicated by the characters RT.

Favorite: Favorites are used by users when they like a tweet. By favoriting a tweet a user can let the original poster know that you liked their tweet. The total number of times a tweet has been favorite is visible to everyone.

Verified User: A verified Twitter user is a user that Twitter has confirmed to be the real. Verification is done by Twitter to establish authenticity of identities of key individuals and brands. The verified status of a user is visible to everyone.

Followers: The followers of a user are other people who receive the user's tweets and updates. When a user is followed by someone, it will show up in their followers list. The total number of followers a user has is visible to everyone.

Followees: The followees of a user are other people who the user followers. The total number of followees a user has is also visible to everyone.

Follower Graph: The graph of users on Twitter and their follower relationship. The nodes in the graph are users and the directional edges represent follower relationship between users.

2. System Design and Architecture:

2.1 Overall 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. 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. On Twitter, a rumor is a collection of tweets, all asserting the same unverified statement (however the tweets could be, and almost assuredly, are worded differently from each other), propagating through the communications network (in this case Twitter), in a multitude of cascades. A rumor can end in three ways: it can be resolved as either true (factual), false (nonfactual) or remain unresolved. 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. 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.

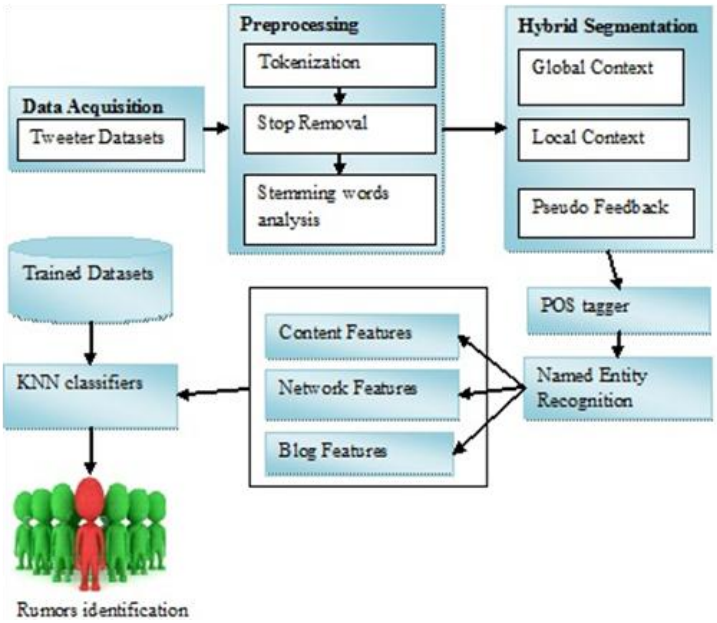


Figure 1: Overall System Architecture

3. Literature Survey:

Twi NER: Named Entity Recognition in Targeted Twitter Stream. This paper presents a novel unsupervised NER system for targeted tweet streams, called Twi NER. Based on the gregarious property of named entities in targeted tweet stream, Twi NER recognizes named entities collectively from a batch of tweets in unsupervised manner. More formally, let T be the collection of tweets in question. Twi NER receives tweets from T in a batch manner. A batch is the set of tweets posted in the targeted Twitter stream within one fixed time interval (e.g. a second). It is noted that currently Twi NER does not categorize the type of named entity (e.g., person, location). As conventional NER methods fail to address the new challenges posed by emerging social media like Twitter, it is more pressing to be able to discover the presence of named entities in targeted Twitter stream before we could categorize their types. Furthermore, even without categorizing the types of named entities, Twi NER already enable us to make early crisis response. For example, a cosmetic company may be interested in discovering any new named entity which may directly/ indirectly link to the company and subsequently causes a PR crisis, be it a person name, product name, or company name. Moreover, as a targeted Twitter stream is constructed for a particular information need, we assume that the user who constructs the stream has the background knowledge in interpreting the named entities detected. Twi NER combines both ideas in a dynamic programming algorithm to efficiently test various segmentation combinations. Note that in this step, we do not use any local linguistic features of a segment, such as its capitalization. Instead, we leverage on the World Wide Web to derive the segmentation.

Algorithm: Twi NER approach

Advantages: Present a novel 2-step unsupervised NER system for targeted Twitter stream, called Twi NER. Constructs a random walk model to exploit the gregarious property in the local context derived from the Twitter stream.

Disadvantages: Difficult to recognize presence of named entities in tweets

Re-ranking for Joint Named-Entity Recognition and Linking Much of the previous research on entity linking has gone into improving linking accuracy over gold-standard mentions, but we observe that many of the common errors made by entity linkers in practice have to do with the pipeline architecture, which propagates errors from named-entity recognition systems to the entity linkers. We introduce a re-ranking model that performs joint named entity recognition and entity linking. The discriminative re-ranking framework allows us to introduce features into the model that capture the dependency between entity linking decisions and mention boundary decisions, which existing models do not handle. Furthermore, the model can handle collective classification of entity links, at least for nearby groups of entities. The joint NER and EL model has strong empirical results, outperforming a number of state-of the art NER and EL systems on several benchmark datasets while remaining computationally inexpensive. In contrast, our techniques are better suited for longer documents. We use linear maximum entropy models to re-rank a set of candidate mentions and entities provided by efficient NER and EL base models. A more minor difference is that Guo et al. link to Wikipedia; our technique links to both Wikipedia and Freebase—a large, user-contributed, relational database. Also, Guo et al.'s techniques cannot identify mention boundaries that have no corresponding Wikipedia entries, whereas our techniques can identify mentions with no corresponding entity in our reference set; we follow the Text Analysis Conference's (TAC) guidelines in linking such mentions to the special symbol NIL.

Algorithm: Entity Linking (EL) system

Advantages: Present a joint model for NER and EL, called NEREL, that takes a large set of candidate mentions from typical NER systems and a large set of candidate entity links from EL systems.

Disadvantages: There is no dependency between entity linking decisions and mention boundary decisions.

Emoticon Smoothed Language Models for Twitter Sentiment Analysis Sentiment analysis (SA) (also known as opinion mining) is mainly about discovering "what others think" from data such as product reviews and news articles. On one hand, consumers can seek advices about a product to make informed decisions in the consuming process. On the other hand, vendors are paying more and more attention to online opinions about their products and services. Hence, SA has attracted increasing attention from many research communities such as machine learning, data mining, and natural language processing. The sentiment of a document or sentence can be positive, negative or neutral. Hence, SA is actually a three-way classification problem. In practice, most methods adopt a two-step strategy for SA. In the subjectivity classification step, the target is classified to be subjective or neutral (objective), and in the polarity classification step, the subjective targets are further classified as positive or negative. Hence, two classifiers are trained for the whole SA process, one is called subjectivity classifier, and the other is called polarity classifier. Since formulated SA as machine learning based text classification problem, more and more machine learning methods have been proposed for SA. As for the models with noisy labels, it is hard for them to achieve satisfactory performance due to the noise in the labels although it is easy to get a large amount of data for training. Hence, the best strategy is to utilize both manually

labeled data and noisy labeled data for training. However, how to seamlessly integrate these two different kinds of data into the same learning framework is still a challenge. In this paper, we present a novel model, called emoticon smoothed language model (ESLAM), to handle this challenge. The basic idea is to train a language model based on the manually labeled data, and then use the noisy emoticon data for smoothing.
Algorithm: Emoticon smoothed language model (ESLAM)
Advantages: Present a novel Emoticon smoothed language model (ESLAM) model, called emoticon smoothed language model (ESLAM) to train a language model based on the manually labeled data, and then use the noisy emoticon data for smoothing.
Disadvantages: Difficult to analyze noise labels

Exploiting Hybrid Contexts for Tweet Segmentation; In this paper, we propose a hybrid tweet segmentation framework incorporating local contexts into the existing external knowledge bases, and name our method Hybrid Seg. Hybrid Seg conducts tweet segmentation in batch mode. Following the same scope of, we only segment tweets from a targeted Twitter stream. A targeted Twitter stream receives tweets based on user defined criteria (e.g., tweets containing some predefined hash-tags or keywords, tweets published by a predefined list of users, or tweets published by users from a specific geographical region). Tweets from a targeted Twitter stream are grouped into batches by their publication time using a fixed time interval (e.g., an hour or a day). Each batch of tweets is then segmented by Hybrid Seg collectively. Hybrid Seg conducts tweet segmentation in an iterative manner. At the first iteration, Hybrid Seg segments the tweet by utilizing the local linguistic features of the tweet itself. To avoid implementation from scratch, we simply apply a set of existing NER tools trained over general English languages on tweets. These existing NER tools provide an initial collection of confident segments by voting. Initializing Hybrid Seg with a set of off-the-shelf NER tools is based on the observation that some tweets from official accounts of news agencies, organizations, advertisers, and celebrities are likely well written. A small set of named entities extracted from these tweets based on voting of classic NER tools can be a high precise yet low recall solution of tweet segmentation.
Algorithm: Hybrid segmentation
Advantages: Exploits the local linguistic features in a collective manner by using the existing NER tools. The recognized named entities with high confidence positively enhance the performance of tweet segmentation.
Disadvantages: Local word dependency is difficult to predict

Recognizing Named Entities in Tweets; In this paper propose a novel NER system to address these challenges. Firstly, a K-Nearest Neighbors (KNN) based classifier is adopted to conduct word level classification, leveraging the similar and recently labeled tweets. Following the two-stage prediction aggregation methods, such pre-labeled results, together with other conventional features used by the state-of-the-art NER systems, are fed into a linear Conditional Random Fields (CRF) model, which conducts fine-grained tweet level NER. Furthermore, the KNN and CRF model are repeatedly retrained with an incrementally augmented training set, into which high confidently labeled tweets are added. Indeed, it is the combination of KNN and CRF under a semi-supervised learning framework that differentiates ours from the existing. The underlying idea of our method is to combine global evidence from KNN and the gazetteers with local contextual information, and to use common knowledge and unlabeled tweets to make up for the lack of training data. We propose to a novel method that combines a KNN classifier with a conventional CRF based labeler under a semi-supervised learning framework to combat the lack of information in tweet and the

unavailability of training data. And evaluate our method on a human annotated data set, and show that our method outperforms the baselines and that both the combination with KNN and the semi-supervised learning strategy are effective.
Algorithm: K-Nearest Neighbors (KNN) classifier
Advantages: Adopted to conduct word level classification, leveraging the similar and recently labeled tweets
Disadvantages: Provide negative effects for learning

A Fast Decoder for Joint Word Segmentation and POS-Tagging Using a Single Discriminative Model; In this paper we follow the line of single-model research, in particular the global linear model of Z&C08. We show that effective decoding can be achieved with standard beam-search, which gives significant speed improvements compared to the decoding algorithm of Z&C08, and achieves accuracies that are competitive with the state-of-the-art. The research is also in line with recent research on improving the speed of NLP systems with little or no accuracy loss. 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, it is combined with existing candidates in all possible ways to generate new candidates, and an agenda is used to keep the N-best candidate outputs from the beginning of the sentence to the current character. Compared to the multiple-beam search algorithm of Z&C08, the use of a single beam can lead to an order of magnitude faster decoding speed. In this paper, take a different approach, and assign a POS-tag to a partial word when its first character is separated from the final character of the previous word. When more characters are appended to a partial word, the POS is not changed. 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, and the likely ones will be kept in the beam for the next processing steps.
Algorithm: Single-beam decoder
Advantages: Can be effectively applied to the decoding problem for a global linear model for joint word segmentation and POS-tagging
Disadvantages: Little accuracy for tweet segmentation

4. Implementation:

4.1 Tweets Acquisition:

Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". Registered users can read and post tweets, but unregistered users can only read them. Users access Twitter through the website interface, SMS, or mobile device app. In order to have an opinion about the user, his posts have to be examined. Therefore, using Twitter API, all tweets posted by user are crawled first. In this study, we tried to examine the user with not only his posts but also his friends' posts. However, crawling all friends' posts is a huge overload, and misleading since Twitter following mechanism does not show an actual interest every time. People sometimes tend to follow some users for a temporary occasion and then forget to un-follow. Sometimes they follow some users just to be informed of, although they are not actually interested in. There are also friends that do not post a tweet for a long time, but still followed by the user. In this module, we can upload the tweet datasets as CSV file. It contains following id, followers id, time stamp, user following, user followers and tweets. In computing, comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

Tweets are publicly visible by default, but senders can restrict message delivery to just their followers. Users can tweet via the Twitter website, compatible external applications (such as for smart phones), or by Short Message Service (SMS) available in certain countries. Users may subscribe to other users' tweets—this is known as "following" and subscribers are known as "followers" or "tweets", a portmanteau of Twitter and peeps. Individual tweets can be forwarded by other users to their own feed, a process known as a "retweet". Users can also "like" (formerly "favorite") individual tweets. Twitter allows users to update their profile via their mobile phone either by text messaging or by apps released for certain smart phones and tablets. Twitter has been compared to a web-based Internet Relay Chat (IRC) client.[118] In a 2009 Time essay, technology author Steven Johnson described the basic mechanics of Twitter as "remarkably simple"

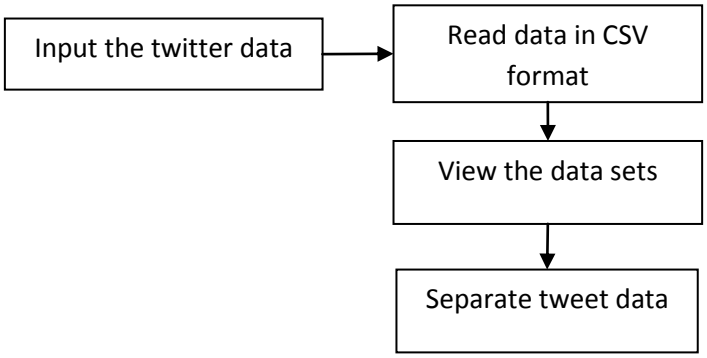


Figure 2: Separate Tweet Data

4.2 Preprocessing:

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. For named entities to be extracted successfully, the informal writing style in tweets has to be handled. Before real data has entered our lives, studies on the area were being conducted on formal texts such as news articles. Generally named entities are assumed as words written in uppercase or mixed case phrases where uppercased letters are at the beginning and ending, and almost all of the studies bases on this assumption. However, capitalization is not a strong indicator in tweet-like informal texts, sometimes even misleading. As the example of capitalization shows, the approaches have to be changed. To extract named entities in tweets, the effect of the informality of the tweets has to be minimized as possible. To obtain this minimalism, following tasks are applied on the data:

- ✓ Links, hash tags, and mentions are removed since they cannot be a part of a named entity.
- ✓ Conjunctives, stop words, vocatives, and slang words etc. are removed.

- ✓ Although punctuation is not taken as an indicator since tweets are informal, still elimination of punctuation is needed. So, smileys are also removed.
- ✓ Repeating characters to express feelings are removed.
- ✓ Informal writing style related issues such as mistyping are corrected.
- ✓ Acidifications related problems are solved since users connecting from mobile devices tend to ignore Turkish characters.

Text mining is an emerging technology that can be used to augment existing data in corporate databases by making unstructured text data available for analysis. An excellent introduction to text mining is provided and provides a short introduction to text mining with a focus on insurance applications. One of the difficulties in getting started with text mining is acquiring the tools, i.e., the software for implementing text mining. Much of the software is expensive and/or difficult to use. For instance, some of the software requires purchase of expensive data mining suites. Other commercial software is more suited to large scale industrial strength applications such as cataloging and searching academic papers. One innovation that has made budget-friendly software available to text miners is open source software. Text mining is a programming language that has wide acceptance for statistical and applied mathematical applications. Perl is a programming language that is particularly strong in text processing. Both languages can be easily downloaded from the Internet. Moreover, a large body of literature now exists to guide users through specialized applications such as text mining.

It can be seen that preprocessing tasks can be divided into two logical groups. Pre-segmenting, and Correcting. Removal of links, hash-tags, mentions, conjunctives, stop words, vocatives, slang words and elimination of punctuation are considered as pre segmentation. It is accepted that parts in the texts before and after a redundant word, or a punctuation mark cannot form a named entity together, therefore every removal of words is behaved as it segments the tweet as well as punctuation does it naturally. Since tweets are pre-segmented before they are handled in tweet segmentation process, pre-segmentation tasks reduces the complexity of the text and increase.

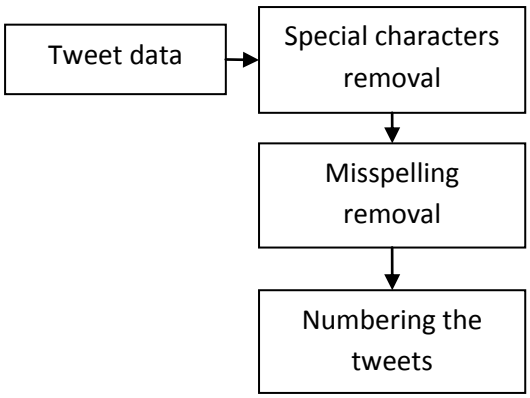


Figure 3: Preprocessing

4.3 Hybrid Segmentation:

Hybrid Seg learns from both global and local contexts, and has the ability of learning from pseudo feedback. Hybrid Seg is also designed to iteratively learn from confident segments as pseudo feedback. Tweets are posted for information sharing and communication. The named entities and semantic phrases are well preserved in tweets. The global context derived from Web pages therefore helps identifying the meaningful segments in tweets. The well preserved linguistic features in these tweets facilitate

named entity recognition with high accuracy. Each named entity is a valid segment. The method utilizing local linguistic features is denoted by Hybrid Seg NER. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge, denoted by Hybrid Seg NGram, is proposed based on the observation that many tweets published within a short time period are about the same topic. Hybrid Seg NGram segments tweets by estimating the term-dependency within a batch of tweets. The segments recognized based on local context with high confidence serve as good feedback to extract more meaningful segments. The learning from pseudo feedback is conducted iteratively and the method implementing the iterative learning is named Hybrid Seg Iter.

In local context, we can eliminate stop words and stemming words based on POS tagger. In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc. And eliminate the stop words from tweets. In computing, stop words are words which are filtered out before or after processing of natural language data (text). Though stop words usually refer to the most common words in a language, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools specifically avoid removing these stop words to support phrase search. Any group of words can be chosen as the stop words for a given purpose. For some search engines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as "The Who", "The The", or "Take That". Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem.

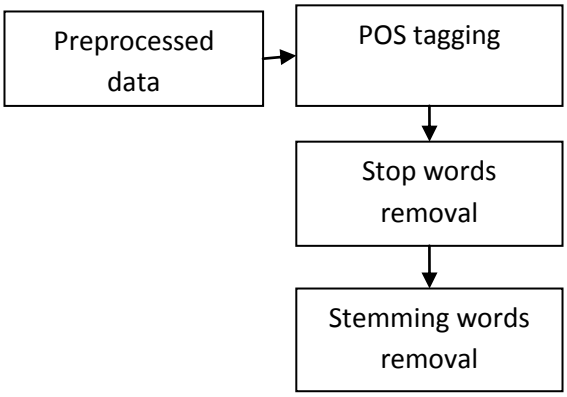


Figure 4: Hybrid Segmentation

4.4 Named Entity Recognition:

Named Entity Recognition can be basically defined as identifying and categorizing certain type of data (i.e. person, location, organization names, and date-time and numeric expressions) in a certain type of text. On the other hand, tweets are characteristically short and noisy. Given the limited length of a tweet, and restriction

free writing style, named entity recognition on this type of data become challenging. After basic segmentation, a great number of named entities in the text, such as personal names, location names and organization names, are not yet segmented and recognized properly. Part of speech tagging is applicable to a wide range of NLP tasks including named entity segmentation and information extraction. Named Entity Recognition strategies vary on basically three factors: Language, textual genre and domain, and entity type. Language is very important because language characteristics affect approaches. Assign each word to its most frequent tag and assign each Out of Vocabulary (OOV) word the most common POS tag. Textual genre is another concept whose effects cannot be neglected. In corpus linguistics, part-of-speech tagging (POS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

Once performed by hand, POS tagging is now done in the context of computational linguistics, using algorithms which associate discrete terms, as well as hidden parts of speech, in accordance with a set of descriptive tags. POS-tagging algorithms fall into two distinctive groups: rule-based and stochastic. E. Brill's tagger, one of the first and most widely used English POS-taggers, employs rule-based algorithms.

Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken. This is not rare in natural languages (as opposed to many artificial languages), a large percentage of word-forms are ambiguous. For example, even "dogs", which is usually thought of as just a plural noun, can also be a verb: The sailor dogs the hatch. Correct grammatical tagging will reflect that "dogs" is here used as a verb, not as the more common plural noun. Grammatical context is one way to determine this; semantic analysis can also be used to infer that "sailor" and "hatch" implicate "dogs" as 1) in the nautical context and 2) an action applied to the object "hatch" (in this context, "dogs" is a nautical term meaning "fastens (a watertight door) securely").

In this module eliminate the rumors using KNN classification. K-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors

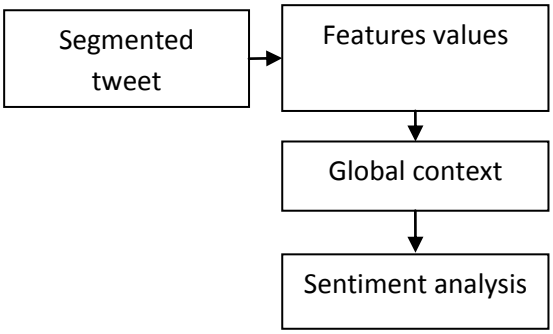


Figure 5: Named Entity Recognition

4.5 Rumor Identification:

So that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A shortcoming of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm has nothing to do with and is not to be confused with k-means, another popular machine learning technique. The training examples are 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 (a query or test point) 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.

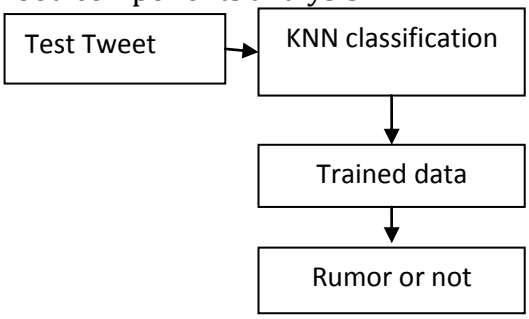


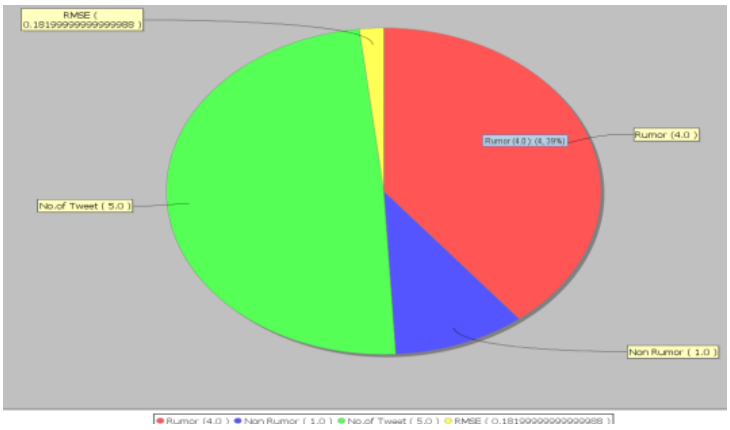
Figure 6: Rumor identification

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:D_X^K
Assign to X the most frequent class in D_X^K
End

5. Experimental Results:



A good strategy of choosing the suitable *K* value in various scenarios is planned for future work. Nevertheless, it is observed that the choice of *K* value depends on the nature of targeted tweet streams, such as topic cohesiveness, *gregarious* property, and size of the tweet collections. Based on the extensive experiments conducted above, we see that by incorporating the encoded intelligence of World Wide Web and local context of tweets. It provides an unsupervised approach for named entity recognition for Twitter, especially for targeted tweet streams with high *gregarious* property.

6. Conclusion:

We designed novel features for use in the classification of tweets in order to develop a system through which informational data may be filtered from the conversations, which are not of much value in the context of searching for immediate information for relief efforts or bystanders to utilize in order to minimize damages. The results of our experiments show that classifying tweets as “rumor” vs. “non rumor” can use solely the proposed features if computing resources are concerned, since the computing power required to process data into featured is immensely decreased in comparison to a BOW feature set which contains a substantially larger number of features. However, if computing power and time necessary to process incoming Twitter data are not a concern, a combined feature set of the proposed features and BOW-presence approach will maximize overall accuracy.

7. Future Enhancement:

In future work, we can extend our approach implement various classification algorithm to predict the attackers and also eliminate the attackers from twitter datasets. And try this approach to implement in various languages in twitter.

8. References:

1. X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In Proc. of ACL, 2011.
2. L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In Proc. of CoNLL, 2009.
3. A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In Proc. of EMNLP, 2011.
4. E. Kouloumpis, T. Wilson, and J. Moore. Twitter Sentiment Analysis: The Good the Bad and the OMG!. ICWSM, 11:538-541, 2011.
5. S. Li, C. Huang, G. Zhou, and S. Lee. Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification. In Proceeding of ACL-10, pp.414-423, 2010.
6. 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.
7. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135.