Indian J.Sci.Res. 14 (1): 102-108, 2017 ISSN: 0976-2876 (Print)
ISSN: 2250-0138 (Online)

# TWEET SEGMENTATION AND CLASSIFICATION FOR RUMOR IDENTIFICATION USING KNN APPROACH

**V. GAYATHRI[a1] AND A.E. NARAYANAN[b]**

[ab]Periyar Maniammai University, Vallam, India

## 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 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 the goal, HybridSeg is generated via named entities extracted from user's followers' and user's own posts. And extend our approach to analyze short text in tweets and rumor based tweets. So implement K-nearest neighbor classifier (K-NN) approach to eliminate rumor based tweets with improved accuracy rates. And can implement in real time tweet environments to identify the rumor with high level security.**

**KEYWORDS:** Social Network, HybridSeg, NER, POS, Data

## BIG DATA

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on." The limitations also affect search finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency,identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5exabytes ($2.5\times10^{18}$) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many peta bytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.

In a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data. In 2012,Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new V "Veracity" is added by some organizations to describe it.

[1]Corresponding author

If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence, regarding data and their use:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviors.

Big data can also be defined as "Big data is a large volume unstructured data which cannot be handled by standard database management systems like DBMS, RDBMS or ORDBMS".

Big data can be described by the following characteristics:

Volume – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

Variability - This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights.


## BIG DATA ADVANTAGES

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. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.

The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. Some people exclusively associate big data with semi-structured and unstructured data of that sort, but consulting firms like Gartner Inc. and Forrester Research Inc. also consider transactions and other structured data to be valid components of big data analytics applications.

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. Mainstream BI software and data visualization tools can also play a role in the analysis process. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a

newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pigas well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems. In some cases, Hadoop clusters and NoSQL systems are being used as landing pads and staging areas for data before it gets loaded into a data warehouse for analysis, often in a summarized form that is more conducive to relational structures. Increasingly though, big data vendors are pushing the concept of a Hadoop data lake that serves as the central repository for an organization's incoming streams of raw data. In such architectures, subsets of the data can then be filtered for analysis in data warehouses and analytical databases, or it can be analyzed directly in Hadoop using batch query tools, stream processing software and SQL on Hadoop technologies that run interactive, ad hoc queries written in SQL. Potential pitfalls that can trip up organizations on big data analytics initiatives include a lack of internal analytics skills and the high cost of hiring experienced analytics professionals. The amount of information that's typically involved, and its variety, can also cause data management headaches, including data quality and consistency issues. In addition, integrating Hadoop systems and data warehouses can be a challenge, although various vendors now offer software connectors between Hadoop and relational databases, as well as other data integration tools with big data capabilities. With big data analytics, scientists and others can analyze huge volumes of data that old analytics and business intelligence solutions can't find. Consider this; it's possible that enterprise could accumulate billions of rows of data with hundreds of millions of data combinations in multiple data stores and abundant formats. Figure is demonstrating the value of Big Data Analytics by drawing the graph between time and cumulative cash flow. Old analytics techniques like any data warehousing application, you have to wait hours or days to get information as compared to Big Data Analytics. Information has the timeliness value when it is processed at right time otherwise it would be of no use. It might not return its value at proper cost.

## EXISTING SYSTEM

### Description

Named entity recognition (NER) is a task of finding and classifying names of things, such as person, location, and organization, given a sequence of words. NER is a very important subtask of information extraction (IE). With the development of the Internet, a huge amount of information has been generated by users. The information generated on the Internet, particularly on social media (e.g., Twitter and Facebook), includes very diverse and noisy texts. Word segmentation is one of the important steps in natural language processing. Essentially, segmentation is trying to determine the boundary of the word. As a fundamental natural language analysis task, word segmentation plays a key role in many natural language processing applications. Different from the traditional word segmentation, many new words exist in the segmentation on twitter. Traditional methods can't deal with this problem well. The need to segment and label sequences arises in many different problems in several scientific fields. Hidden Markov models (HMMs) and stochastic grammars are well understood and widely used probabilistic models for such problems. These approaches difficult to segments each entities in tweets. Twitter is a type of micro blogging service in which users are allowed to post contents such as small messages, individual images, or videos. Base features include the gazetteer features and orthographic features. In the NER task, a huge amount of unlabeled data is often used for identifying unseen entities. There are already 53 gazetteers in the baseline system. The maximum window size for gazetteer features is 6, and the model will learn the named entity type associated with a specific phrase, if it is in one or more of the gazetteer lexicons. Orthographic features can be divided into five types. The orthographic feature templates are as follows:

- n-gram: wi for i in {-1,0,1}, conjunction of previous word and current word wi−1|wi for i in {-1,0}.
- Affixes: Prefixes and suffixes of xi . The first and last n characters ranging from 1 to 3.
- Capitalization: There are two patterns of capitalization: One is an indicator of capitalization for the first character, and the other is an indicator of capitalization for all characters.
- Digit: There are three patterns for numbers: i) Whether the current word has a digit, ii) whether the current word is a single digit, and iii) whether the current word has two digits.
- Non-alphabet: Whether the current word contains a hyphen and other punctuation marks. Among the other punctuation marks is the colon (:). In general, what follows right after a colon mark represents a feature weight. To make the model learn correctly, we normalize only the colon mark

In the NER task, POS tags and chunks contain very useful information for finding and classifying named entities. We predict POS tags and chunks by using a model trained with Twitter data. Commonly used NER methods on formal texts such as newspaper articles are built upon on linguistic features extracted locally. However, considering the short and noisy nature of tweets, performance of these methods is inadequate on tweets and new approaches have to be generated to deal with this type of data. Recently, tweet representation based on segments in order to extract named entities has

proven its validity in NER field. Along with named entities extracted from tweets via tweet segmentation, user's retweet and mention history, and followed users are also considered as strong indicators of interest and a model representing user interest is generated. Reducing Twitter users' effort to access tweets carrying the information of interest is the main goal of the study, and a tweet recommendation approach under a user interest model generated via named entities is presented.

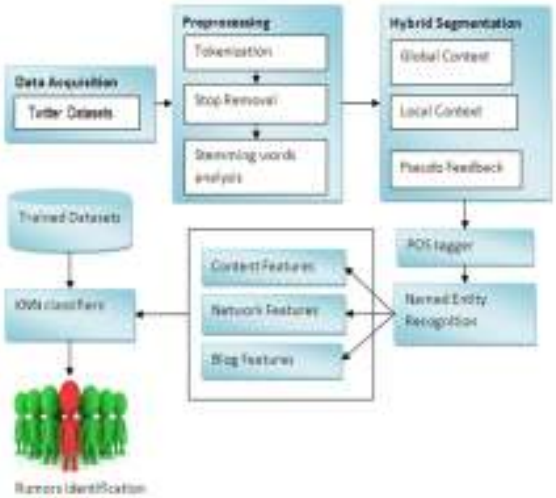
**Disadvantages of the existing system**

- Event detection and summarization, opinion mining, sentiment analysis, and many others.
- Limited length of a tweet (i.e., 140 characters) and no restrictions on its writing styles, tweets often contain grammatical errors, misspellings, and informal abbreviations.

On the other hand, despite the noisy nature of tweets, the core semantic information is well preserved in tweets in the form of named entities or semantic phrases.

## PROPOSED SYSTEM

**System architecture**




**Explanation**

A rumor is defined as a statement whose true value is unverifiable. Rumors may spread misinformation on a social media of people. Identifying rumors is critical in online social media where huge amounts of information are easily reached across a large network by sources with unverified authority. In this paper, we address the problem of rumor detection on twitter a social media. This paper focuses on the development of HybridSeg and KNN approach to the classification of tweets (posts on Twitter). HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback. In order to analyze the textual content of the tweets, we give an overview of the top terms occurring in each type of topic, i.e., the most frequent vocabulary used in each type of topic. To do so, we first performed a filtering process to remove irrelevant words. The filtering process removed all the stop words contained in the tweets. The stop word removal process includes Twitter-specific words and words in stop word lists for the main languages in the dataset. After that, we computed the TF (term frequency) of each word for each type of trending topic. This process produced a list of words for each type of trending topic, ranked in a descending order by TF value. These steps are implemented in Global and Local Context. Before we extract pseudo feedback, POS tagger implements to define features categories such adverb, adjective and so on in Natural Language. Then implement KNN approach to classify the rumors using three features such as content features, network features, blog features.

**Implementation**

- Tweets acquisition
- Preprocessing
- Hybrid segmentation
- Named entity recognition
- Performance evaluation

**Modules Description**

**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

**Preprocessing**

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.
- Artification related problems are solved since users connecting from mobile devices tend to ignore Turkish characters.

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

**Hybrid segmentation**

HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback. HybridSeg 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 HybridSegNER. It obtains confident segments based on the voting results of multiple off-the-shelf NER tools. Another method utilizing local collocation knowledge, denoted by HybridSegNGram, is proposed based on the observation that many tweets published within a short time period are about the same topic. HybridSegNGram 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 HybridSegIter.

**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.

**Performance evaluation**

In this module, we can evaluate the process of the system using accuracy rate and normalized utility. Our proposed system provides improved accuracy rate and normalized utility.

**Design/Algorithm/Pseudo code (whichever is applicable) KNN classification**

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its $k$ nearest neighbors ($k$ is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.

- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

*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, 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.

A commonly used distance metric for continuous variables is Euclidean distance. For discrete variables, such as for text classification, another metric can be used, such as the overlap metric (or Hamming distance). In the context of gene expression microarray data, for example, *k*-NN has also been employed with correlation coefficients such as Pearson and Spearman. 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.

**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)*

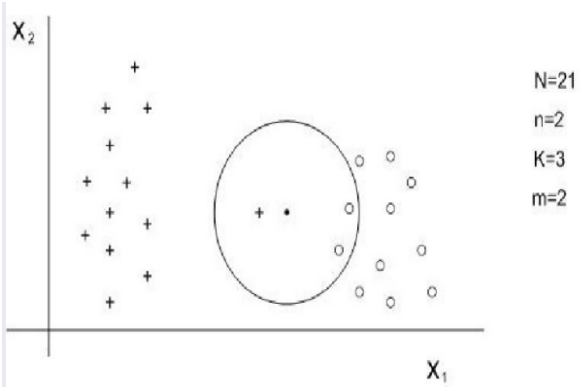
*Order d(Xi,X) from lowest to highest, (i-1,…,N)*

*Select the K nearest instances to X:* $D_x^K$

*Assign to X the most frequent class in* $D_x^K$

*End*

The basic concept as: (3-NN)




## 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.

## FUTURE WORK

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.

## REFERENCES

Li C., Weng J., He Q., Yao Y., Datta A., Sun A. and Lee B.S., "Twiner: Named entity recognition in targeted twitter stream," in SIGIR, 2012, pp. 721–730.

Li C., Weng J., He Q. and Sun A., "Exploiting hybrid contexts for tweet segmentation," in SIGIR, 2013, pp. 523–532.

Liu X., Zhang S., Wei F. and Zhou M., "Recognizing named entities in tweets," in ACL, 2011, pp. 359–367.

Liu X., Zhou X., Fu Z., Wei F. and Zhou M., "Exacting social events for tweets using a factor graph," in AAAI, 2012.

Ritter A., Mausam, Etzioni O. and Clark S., "Open domain event extraction from twitter," in KDD, 2012, pp. 1104–1112.

Ritter A., Clark S., Mausam and O. Etzioni, "Named entity recognition in tweets: An experimental study," in EMNLP, 2011, pp. 1524–1534.

Cui A., Zhang M., Liu Y., Ma S. and Zhang K., "Discover breaking events with popular hashtags in twitter," in CIKM, 2012, pp. 1794–1798.

Meng X., Wei F., Liu X., Zhou M., Li S. and Wang H., "Entity centric topic-oriented opinion summarization in twitter," in KDD, 2012, pp. 379–387.

Luo Z., Osborne M. and Wang T., "Opinion retrieval in twitter," in ICWSM, 2012.

Wang X., Wei F., Liu X., Zhou M. and Zhang M., "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in CIKM, 2011, pp. 1031–1040.