TWEET IDENTIFICATION AND CLASSIFICATION FOR ROMOUR IDENTIFICATION USING KNN APPROACH

[*]M.Selvam Amalraj, Assistant *Professor, Department of Computer Science ,*
Bharathiyar college of Engineering and Technology, Karaikal
***A.Akshaya, II year M.Tech computer science,*
Bharathiyar college of Engineering and Technology, Karaikal

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, HybridSeg 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 Support Vector machine 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.

Keywords: ***Twitter Stream, Tweet Segmentation, Named Entity Recognition, KNN.***

1.INTRODUCTION

Social media is an invaluable source of almost any information. Social media opens up access to an "effective and irrepressible real-time mechanism to broadcast information". 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 project develops models for detection of rumors (i.e. unverified information) that propagate on Twitter. Clustering the tweets using hybrid segmentation algorithm which contains local context and global context with various named entities. Detection of rumors about an event is achieved through classifying and clustering

assertions made about that event. Assertions are classified through SVM classifier for Twitter developed for this project.

The classifier is a logistic regression that utilizes a combination of semantic and syntactic features and can identify assertions with improved accuracy. In this annotation task, the human assessor reads through a timeline of tweets to determine which of these are associated with rumours. Without necessarily having prior knowledge of the rumours associated with a given event, we expect that this approach will let us discover new stories. To facilitate the task, we had to deal with two major issues: (i) the number of tweets tends to be large for any given event, and (ii) a tweet does not always provide enough contexts to be able to determine whether it is referring to a rumour. Twitter, a microblogging site, serves as an immediate form of broadcasting information to the world; it is a place where "people digitally converge during disasters". Informative data from sites such as Twitter can either be obtained directly from bystanders of a disaster or "derivative - that is, information in the form of reposts or pointers to information available elsewhere". Both misinformation and conversation can cloud the entirety of data coming in through social media during disasters.

2. LITERETURE SURVEY

2.1" TwiNER: Named Entity Recognition in Targeted Twitter Stream"

This paper presents a novel unsupervised NER system for targeted tweet streams, called TwiNER. Based on the gregarious property of named entities in targeted tweet stream, TwiNER recognizes named entities collectively from a batch of tweets in unsupervised manner. More formally, let T be the collection of tweets in question. TwiNER 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 TwiNER 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, TwiNER already enable us to make early crisis response.

2.2. "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.

2.3 "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.

2.4 "Twevent: Segment-based Event Detection from Tweets"

In this paper propose to extract social events from multiple similar tweets using a factor graph to tackle this issue. The solution is based on the following observation: a social event is likely to appear in multiple tweets, for some of which extraction is easy while for the others it is hard. It is straightforward to extract an interaction type social event involving "Aquino" and "Obama" from the first tweet, owing to the strong indicator of "meet"; while it is hard to extract it from the second, because of the limited context. Intuitively, knowing that there is an interaction type social event involving "Aquino" and "Obama" in the first tweet will encourage us to guess the same event for the second tweet. Note that the idea of utilizing redundancy in tweets has been successfully practiced in some of recent studies on tweets, most of which are based on the two-stage labeling strategy. We propose to extract social events from multiple similar tweets using a graphical model, which exploits redundancy in tweets to make up for the lack of information in a single tweet. Firstly, we use a graphical model, rather than kernel SVM, to simultaneously infer the labels of all event candidates. One advantage of our model is that it allows us to aggregate across tweet information to make up for the lack of information in a single tweet. Secondly, our model adopts only shallow linguistic features, rather than chunking and dependency parsing related features. This is because tweets are short and often informally written, and as a result current natural language processing tools perform badly on tweets. However, two significant differences exist. Firstly, our work concentrates on social event extraction for tweets, rather than entertainment events; secondly, feature weights of our factor graph are automatically tuned on the training data.

3. PROPOSED WORK

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 project 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.

3.1 System Architecture

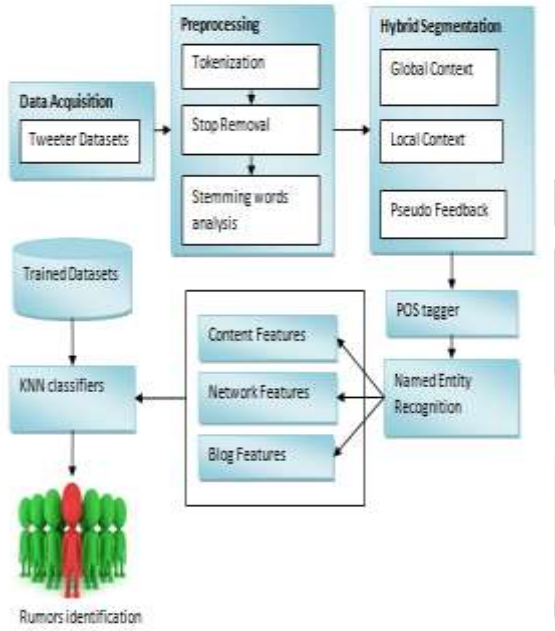

Fig:I System Architecture.

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.

3.2 System Implementation

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in

giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of existing system and its constraints on implementation, design of methods to achieve change over and calculation of change over methods. Implementation is the process of converting a new system design into operation. It is the phase that focuses on user training, site preparation and file conversation for installing a candidate system. The important factor that should be considered here is that the conversation should not disrupt the functioning of the organization.

3.2.1 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.

3.2.2 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.

3.2.3 KNN Classification

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, 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

3.3 Related works

Chenliang Li [1]. This paper presents a novel unsupervised NER system for targeted tweet streams, called TwiNER. Based on the gregarious property of named entities in targeted tweet stream, TwiNER recognizes named entities collectively from a batch of tweets in unsupervised manner. More formally, let T be the collection of tweets in question. TwiNER 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 TwiNER does not categorize the type of named entity (e.g., person, location).

Avirup Sil [2]. 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. we follow the Text Analysis Conference's (TAC) guidelines in linking such mentions to the special symbol NIL.

Kun-Lin Liu, [3]. 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. 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.

Chenliang Li [4]. In this paper propose to extract social events from multiple similar tweets using a factor graph to tackle this issue. The solution is based on the following observation: a social event is likely to appear in multiple tweets, for some of which extraction is easy while for the others it is hard. However, two significant differences exist. Firstly, our work concentrates on social event extraction for tweets, rather than entertainment events; secondly, feature weights of our factor graph are automatically tuned on the training data.

Axis Sun .. [5]. In this paper, we propose a hybrid tweet segmentation framework incorporating local contexts into the existing external knowledge bases, and name our method HybridSeg. HybridSeg conducts tweet segmentation in batch mode. Following the same scope of, we only segment tweets from a targeted Twitter streamHybridSeg conducts tweet segmentation in an iterative manner. At the first iteration, HybridSeg 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 HybridSeg 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.

Kurt Junshean Espinosa [6]. 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. An 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.

Xiaodong Zeng [7]. Extracting useful structured representations of events from this disorganized corpus of noisy text is a challenging problem. On the other hand, individual tweets are short and self-contained and are therefore not composed of complex discourse structure as is the case for texts containing narratives. In this paper we demonstrate that open-domain event extraction from Twitter is indeed feasible, for example our highest-confidence extracted future events are 90% accurate.

Yue Zhang [8]. 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.

4. 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.

5. REFERENCE

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 SIGIR, 2012, pp. 721–730.
[2] C. Li, A. Sun, J. Weng, and Q. He, “Exploiting hybrid contexts for tweet segmentation,” in SIGIR, 2013, pp. 523–532.
[3] A. Ritter, S. Clark, Mausam, and O. Etzioni, “Named entity recognition in tweets: An experimental study,” in EMNLP, 2011, pp. 1524–1534.
[4] X. Liu, S. Zhang, F. Wei, and M. Zhou, “Recognizing named entities in tweets,” in ACL, 2011, pp. 359–367.
[5] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, “Exacting social events for tweets using a factor graph,” in AAAI, 2012.

[6] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, "Discover breaking events with popular hashtags in twitter," in CIKM, 2012, pp. 1794–1798.
[7] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in KDD, 2012, pp. 1104–1112.
[8] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entitycentric topic-oriented opinion summarization in twitter," in KDD, 2012, pp. 379–387.
[9] Z. Luo, M. Osborne, and T. Wang, "Opinion retrieval in twitter," in ICWSM, 2012.