International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Tweet Segmentation and Classification for Rumor Identification using KNN Approach

S.Vinitha[1], Mrs S. Nalini[2]

[1]MCA Student Department of Computer Applications, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu, India

[2]Assistant professor, Department of Computer Applications, Anna University, BIT Campus, Tiruchirappalli, Tamil Nadu, India

ABSTRACT

Twitter is a source of sharing and communicate recent information, ensuing into huge size of records produces every day. Even though, a various applications of Natural Language Processing and Information Retrieval go through rigorously from an erroneous and tiny nature of tweets. We thought to implement a framework in support of segmentation of tweet by collection form, called as HybridSeg. During tweet separating with trivial segments, surroundings information is preserved and simply takes out by the downstream application. HybridSeg glance for top segmentation of a tweet through increasing stickiness score of its candidate segment. The stickiness score is explanation the possibility of a segment is express in English (global context and local context). Finally we advise and assess two models to acquire with local context by concerning the term-dependency in a collection of tweets, in the same way. Testing on two tweet data sets give you an idea about tweet segmentation superiority is considerably enhanced by global and local contexts evaluate by use of global context simply. Assessment and relationship, we demonstrate that additional correctness is accomplished in Named Entity Recognition by part-of-speech (POS).

Keywords: KNN Algorithm , Twitter, Tweet, Tweet segmentation, Named Entity Recognition

I. INTRODUCTION

Twitter, as a latest form of public media, has seen great expansion. It has apprehensive immense interests since both academia and business. Most of private and/or public sector accounted to supervise Twitter stream to collect and understand users' opinions about the combination while, because of large quantity of tweets available daily the nature of tweets obtain a challenge. Normal NER methods on sound-structure documents greatly depend on a phrase's local linguistic features Tweets often consist of informal abbreviations, misspellings in addition to grammatical errors. Tiny nature and an error-prone of tweets repeatedly make the word-level language models for tweets less reliable. Illustration, Recognized a tweet "I call geeta no response her mobile is in the bag, she running", there is no idea regarding its true theme by disregard word sequence. The circumstance is further exacerbated with the inadequate context provided by the tweet .

II. LITERATURE SURVEY

[Biao Wang] With the soaring development of large scale online social networks, online information sharing is becoming ubiquitous everyday. Various information is propagating through online social

networks including both the positive and negative. In this paper, we focus on the negative information problems such as the online rumors. Rumor blocking is a serious problem in large-scale social networks. Malicious rumors could cause chaos in society and hence need to be blocked as soon as possible after being detected. In this paper, we propose a model of dynamic rumor influence minimization with user experience (DRIMUX). Our goal is to minimize the influence of the rumor (i.e., the number of users that have accepted and sent the rumor) by blocking a certain subset of nodes. A dynamic Ising propagation model considering both the global popularity and individual attraction of the rumor is presented based on realistic scenario. In addition, different from existing problems of influence minimization, we take into account the constraint of user experience utility. Specifically, each node is assigned a tolerance time threshold. If the blocking time of each user exceeds that threshold, the utility of the network will decrease. Under this constraint, we then formulate the problem as a network inference problem with survival theory, and propose solutions based on maximum likelihood principle. Experiments are implemented based on large-scale real world networks and validate the effectiveness of our method.

[Xiaohua Liu] This paper presents Social events are events that occur between people where at least one person is aware of the other and of the event taking place. Extracting social events can play an important role in a wide range of applications, such as the construction of social network. In this paper, we introduce the task of social event extraction for tweets, an important source of fresh events. One main challenge is the lack of information in a single tweet, which is rooted in the short and noise-prone nature of tweets. We propose to collectively extract social events from multiple similar tweets using a novel factor graph, to harvest the redundance in tweets, i.e., the repeated occurrences of a social event in several tweets. We evaluate our method on a human annotated data set, and show that it outperforms all baselines, with an absolute gain of 21% in F1. Social events are events that occur between people, in which at least one person is aware of the other and of the event taking place (Agarwal and Rambow 2010). For example, there is a social event in the sentence “John talks to Mary”, since John and Mary are aware of each other and both are aware of the talking event.

[Alan Ritter, Sam Clark, Mausam and Oren Etzioni Computer Science and Engineering] People tweet more than 100 Million times daily, yielding a noisy, informal, but sometimes informative corpus of 140-character messages that mirrors the zeitgeist in an unprecedented manner. The performance of standard NLP tools is severely degraded on tweets. This paper addresses this issue by re-building the NLP pipeline beginning with part-of-speech tagging, through chunking, to named-entity recognition. Our novel T-NER system doubles F1 score compared with the Stanford NER system. T-NER leverages the redundancy inherent in tweets to achieve this performance, using LabeledLDA to exploit Freebase dictionaries as a source of distant supervision. LabeledLDA outperforms cotraining, increasing F1 by 25% over ten common entity types. Status Messages posted on Social Media websites such as Facebook and Twitter present a new and challenging style of text for language technology due to their noisy and informal nature. Like SMS (Kobus et al., 2008), tweets are particularly terse and difficult (See Table 1). Yet tweets provide a unique compilation of information that is more upto- date and inclusive than news articles, due to the low-barrier to tweeting, and the proliferation of mobile devices.1 The corpus of tweets already exceeds.

[Chenliang Liy, Aixin Suny, Jianshu Wengz, Qi Hex ySchool of Computer Engineering, Nanyang Technological University, Singapore] Twitter has attracted hundred millions of users to share and

disseminate most up-to-date information. However, the noisy and short nature of tweets makes many applications in information retrieval (IR) and natural language processing (NLP) challenging. Recently, segment-based tweet representation has demonstrated e_ectiveness in named entity recognition (NER) and event detection from tweet streams. To split tweets into meaningful phrases or segments, the previous work is purely based on external knowledge bases, which ignores the rich local context information embedded in the tweets. In this paper, we propose a novel framework for tweet segmentation in a batch mode, called HybridSeg. HybridSeg incorporates local context knowledge with global knowledge bases for better tweet segmentation. HybridSeg consists of two steps: learning from o_- the-shelf weak NERs and learning from pseudo feedback. In the first step, the existing NER tools are applied to a batch of tweets. The named entities recognized by these NERs are then employed to guide the tweet segmentation process. In the second step, Hybrid- Seg adjusts the tweet segmentation results iteratively by exploiting all segments in the batch of tweets in a collective manner. Experiments on two tweet datasets show that HybridSeg significantly improves tweet segmentation quality compared with the state-ofthe-art algorithm. We also conduct a case study by using tweet segments for the task of named entity recognition from tweets. The experimental results demonstrate that HybridSeg significantly benefits the downstream applications.

[Chenliang Li, School of Computer Engineering, Nanyang Technological University, Singapore] Twitter, as a new type of social media, has seen tremendous growth in recent years. It has attracted great interests from both industry and academia. Many private and/or public organizations have been reported to monitor Twitter stream to collect and understand users' opinions about the organizations. Nevertheless, due to the extremely large volume of tweets published every day1, it is practically infeasible and unnecessary to listen and monitor the whole Twitter stream. Therefore, targeted Twitter streams are usually monitored instead; each such stream contains tweets that potentially satisfy some information needs of the monitoring organization. Targeted Twitter stream is usually constructed by filtering tweets with user-defined selection criteria depends on the information needs. For example, the criterion could be a region so that users' opinions from that particular region are collected and monitored; it could also be one or more predefined keywords so that opinions about some particular events/topics/products/services can be monitored.

GLOBAL KNOWLEDGE GUIDED TWEET SEGMENTATION

Global knowledge bases like Wikipedia and Microsoft Web N-Gram corpus [24] have been successfully utilized to guide the tweet segmentation, which is named GlobalSeg by this paper. In this section, we briefly review its process, as illustrated by Figure 1.The input of GlobalSeg is a tweet d = w1w2 : : :w`, and its output is m (m _ `) non-overlapped segments, d = s1s2 : : : sm, where a segment si is a n-gram (n _ 1). The optimal segmentation is to
maximize the sum of stickiness scores of m segments:
arg max
s1;:::;sm
Xm
i=1
C(si); (1)
where C(s) denotes a stickiness function, defined as:
C(s) = L(s) _ eQ(s) _2 1 + e SCP(s) : (2)
Eq. 2 encodes three factors for measuring the stickiness of each segment. They are: 1) the normalized segment length L(s) = 1
if jsj = 1 and L(s) = jsj 1 jsj if jsj > 1 which moderately alleviates the penalty on long segments; 2) the probability that s is inside an anchor text in Wikipedia Q(s); 3) the Symmetric Conditional Probability (SCP) measure defined by SCP(s) = log

Pr(s)2 1 jsj1
Pjsj1 i=1 Pr(w1 : : :wi) Pr(wi+1 : : :wjsj) ; (3)
Pr(s) is the n-gram probability provided by MicrosoftWeb N-Gram
service. If s is a single word w, SCP(s) = 2 log Pr(w). If s contains
more words, SCP tends to keep a cohesive s while all its possible
binary partitions are not cohesive.

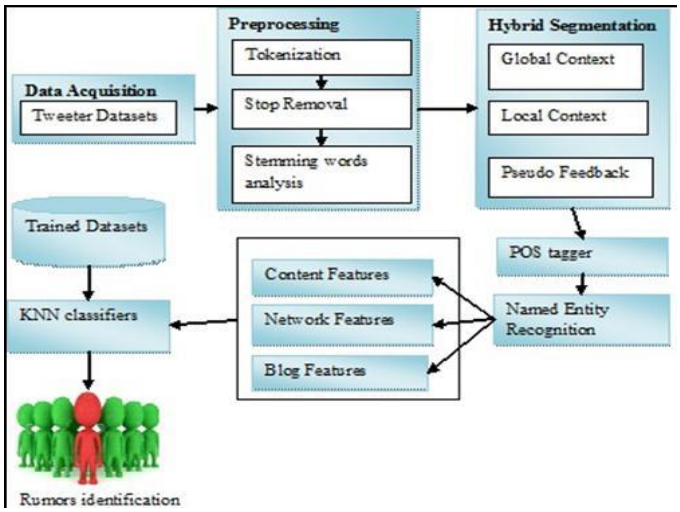
EXISTING SYSTEM

Now a days twitter is a most familiar social media in globe. They can easily understand what they want to say in a single word also avoid large number of data space management system. Just simplified the convertion rate and usage of twitter.

PROPOSED SYSTEM

The proposed system describes the Hybrids Approach and to finds the Optimal Segmentation of Tweets. Hybrids is generated via Named Entities Extracted from User's Followers' and User's Own Posts. It is difficult to classify Rumors in each Tweets, to implement the K- Nearest Neighbor Classifier (K-NN) Approach to Eliminate short text in Rumor Based Tweets.

There are attempt that design linguistic features to capture tweets unique characteristic and train tweet – specific models al trained a POS tagger with the help of a new labeling scheme and afuture set that capture the unique characteristic of tweets .

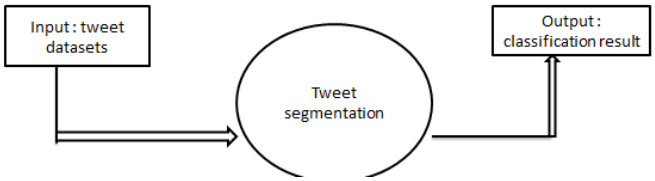


Twitter is a micro-blogging social media platform with hundreds and millions of users. Twitter is a social network where users can publish and exchange short messages of up to 140 characters long, also known as tweets. It can define a rumor to an unverified assertion that starts from one or more sources and spreads over time from node to node in a network. Figure 1.2 explains the architectural design of tweet processing with the big data perspective. The Short text datasets are collected in the data acquisition process, then in Data pre-processing it includes cleaning, normalization, transformation and Stemming words analysis. The keywords are analyzed based on POS tagger. Next, the Hybird segmentation process is done, HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback. The segments recognized based on local context with high confidence serve as good feedback to extract more meaningful segments. 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. Now the KNN approach is designed to classify the short text in rumors based tweets. k-NN algorithm is the simplest of all machine learning algorithms. KNN classification approach is used to label the each tweets. This process eliminate the rumors using KNN classification. Finally, a system that can detect short text message as rumors and

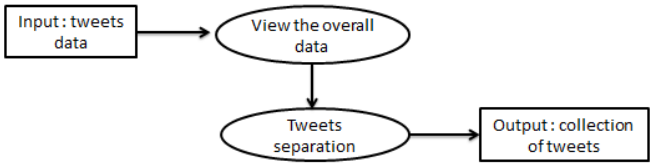
predict their veracity and maybe impact is indeed a very valuable and useful tool.

Data flow diagram

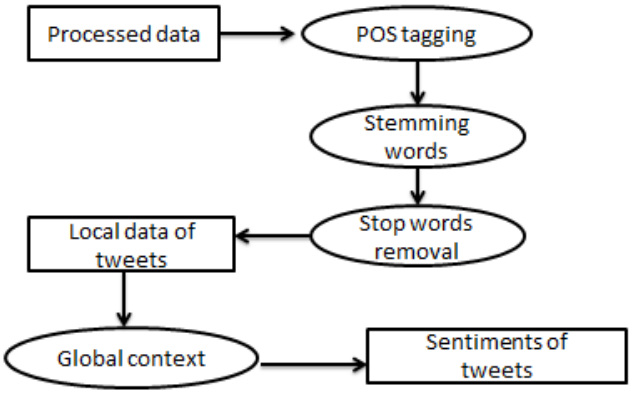
Level 0



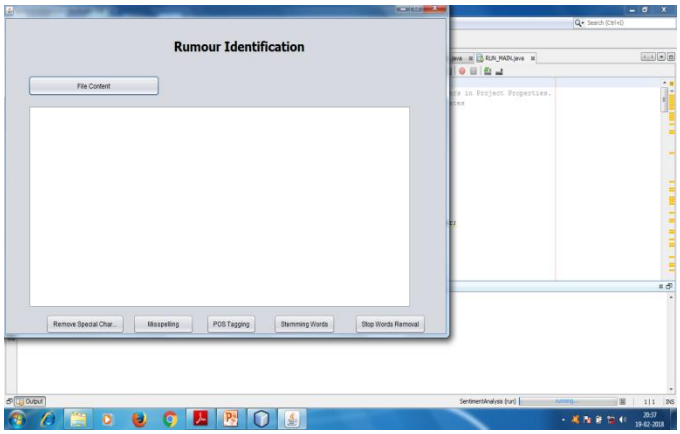
Level 1

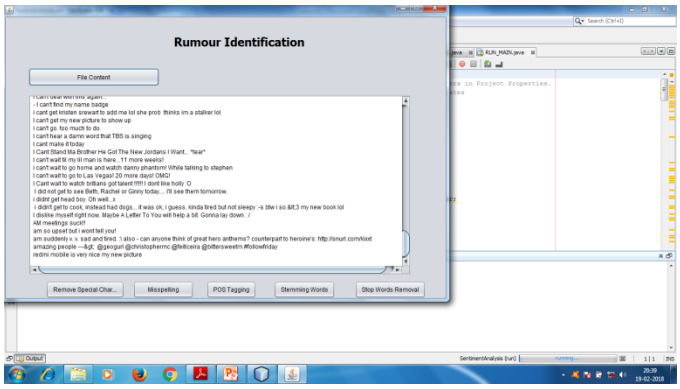


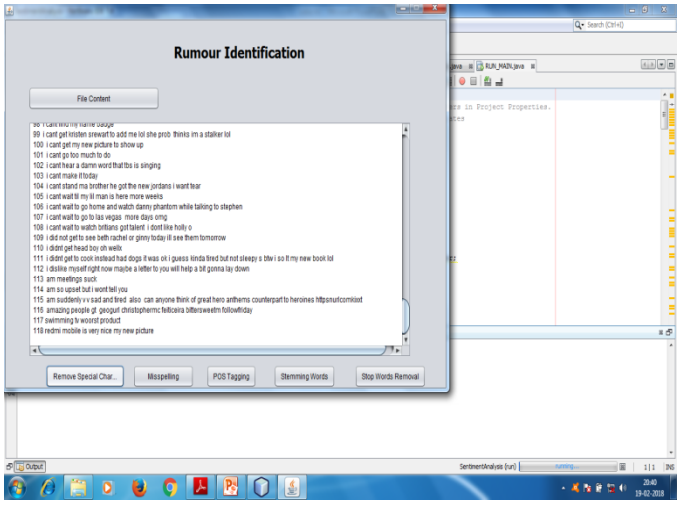
Level 3

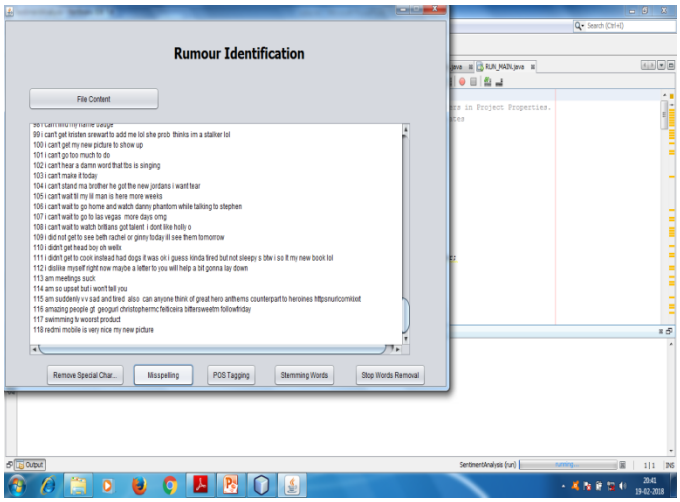


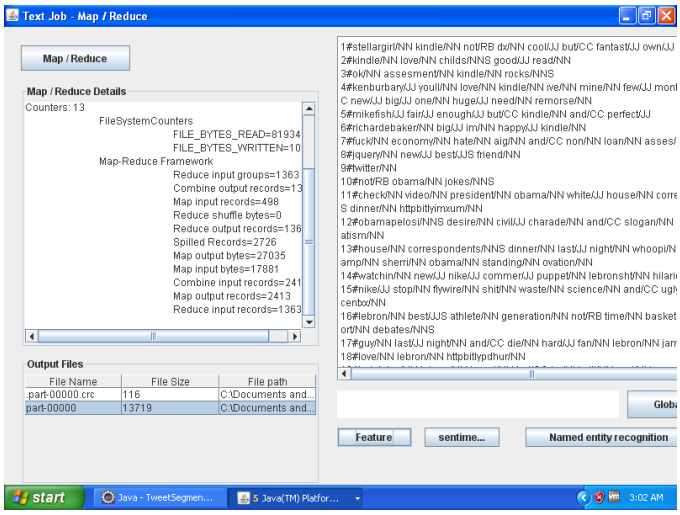
Output :

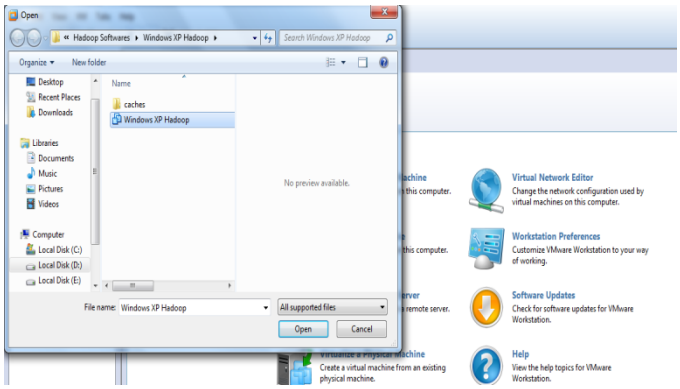


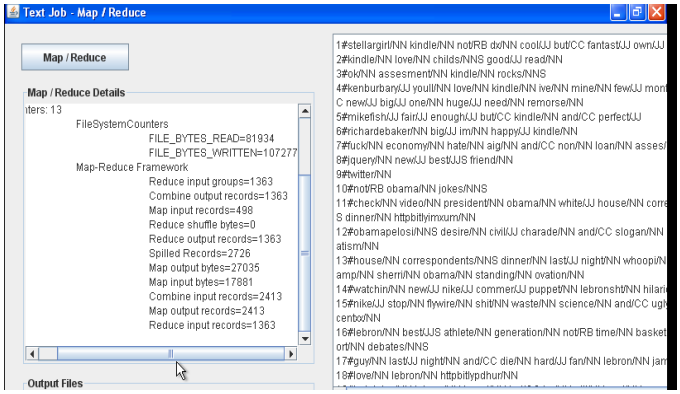


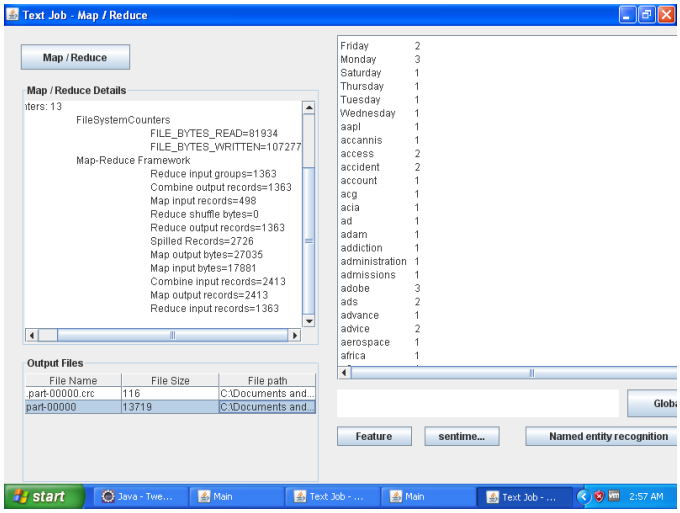












III. CONCLUSION

In this work, we implemented KNN classification algorithms for tweet segmentation, the KNN classification was very effective in tweet segmentation. The main aim is to build a system that employs this work and the emerging patterns within the re-tweet network topology to find whether a short text is rumor or not. This work involves using more advanced techniques from linguistics to extend the speed of correct tense identification. In specific, developments to the analysis of verb phrases and modifications of the marked parameters for sentences might be terribly useful. By creating it simple to match news coverage to twitter posts concerns a happening, the system offers both up-to- the-minute information and valuable insight into past events. In future ,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. market-place. The main goal of this system is for effective data summarization and characterization using visualization techniques.

IV. REFERENCES

1. Doug Downey, Matthew Broadhead, and Oren Etzioni. 2007. Locating complex named entities in web text. In Proceedings of the 20th

international joint conference on Artifical intelligence.

2. K Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for twitter: annotation, features, and experiments. In ACL-HLT, pages 42-47, 20113. 3. M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist., 23(1):33-64, Mar. 1997.
3. K-L. Liu, W.-J. Li, and M. Guo. Emoticon smoothed language models for twitter sentiment analysis. In AAAI,.
4. Mark Hachman. 2011. Humanity's tweets: Just 20 terabytes. In PCMAG.COM.
5. D N. Milne and I. H. Witten, "Learning to link with wikipedia," in CIKM, 2008, pp. 509-518..
6. W Jiang, L. Huang, and Q. Liu, "Automatic adaption of annotation standards: Chinese word segmentation and pos tagging - a case study," in ACL, 2009, pp. 522-530.
7. Han, B., and Baldwin, T. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In ACL, 368-378
8. K Wang, C. Thrasher, E. Viegas, X. Li, and P. Hsu. An overview of microsoft web n-gram corpus and applications. In Proc. of NAACL-HLT, 2010.
9. Y. Wang. Annotating and recognising named entities in clinical notes. In Proc. of the ACL-IJCNLP 2009 Student Research Workshop, 2009.