Volume 5, Issue 11, November 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Online Tweet Summarization for Linguistic Features

Sreeja G. G
PG Scholar,
Dept of CSE, Sri Shakthi Institute of Engineering and Technology, India

Narmadha R. P.
Asst. Prof
Dept. of CSE, Sri Shakthi Institute of Engineering and Technology, India

Abstract- Tweet classification has been used for classifying the tweets based on different class labels contained in the tweets. NLP models have to be constructed in order to learn the tweets with linguistic feature .Before feature extraction of the data; tweets are pre-processed with stop word removal and Stemming process. Initially tweets will be organized into meaningful segments based on Local and global context of the tweet information. Tweet Summarization is proposed in order to avoid the overload problems due to diversity among the sentences, large number of tweets is meaningless, irrelevant and redundant. Further, tweets are strongly correlated with their posted time and new tweets tend to arrive at a very fast rate. Classification technique which establishes the optimal cluster of a tweet is carried in terms of tweet cluster vector. Additionally tweet vector cluster is established as potential sub-topic delegates and maintained dynamically in memory during stream processing. Data structure is used to store and organize cluster snapshots at different moments, thus allowing historical tweet data to be retrieved by any arbitrary time durations.

Keywords: Tweet classification, Pre-processing data, Tweet Summarization, Linguistic features, Feature Extraction.

I. INTRODUCTION

Microblogging services such as Twitter has extracted millions of users to share their information between the people and extract knowledge from the shared information, [2] as they offer large volumes of real time data, with around 200 millions of regular users posting the tweets per day in June 2015. Personally, the people may not know the person who is sharing the information but they may get used with that information. For this, the users must be able to understand the tweets so that they can gain some knowledge and also continue their comments on the same topic. Recognizing the importance of Twitter, many researchers have provided some useful works, which allowed users to extract some information from the tweets. Eventhough, the researchers provide some technique to understand the tweets, there may be many grammatical errors, misspellings or any ill-formed words be available, which leads people to misunderstand the actual semantic of the tweets. There is also a possibility for much diversities among the sentences which may not allow the users to understand the tweets clearly. For instance, [4] "I call her no answer. Her phone in the bag, she dancin". In this example, we actually think that a girl is dancing and her phone was in the bag when someone called her so she was not able to answer the call. But here it is a contradiction. The phrase "She dancing" actually represents the most famous song in Bay Area which was a trendy topic in January 2013. In the existing work, word based Named Entity Recognition (NER) method has been employed for analyzing the semantic of the tweets which is less accurate in terms of interclass similarity and execution time.

To overcome the above challenges, an online summarization of tweets has been analyzed in this paper. An online Summarization is nothing but providing the meaningful tweets in a precise way by understanding the actual comments posted in the twitter. Summarization represents a set of documents by a summary consisting of several sentences [8]. This Summarization is extensively used in content representation, especially when users surf the internet with their mobile devices which have much smaller screens than PC's [8]. This new adopted summarization technique uses segment based Named Entity Recognition (NER) method which provides better accuracy than the traditional summarization technique in terms of interclass similarity and execution time.

In the present study, the tweets are partitioned into different segments by avoiding the supporting words in each sentence using stop word removal method. To perform this method, the data must be preprocessed to analyze the words in the tweets. The Stop word removal method is used mainly to eliminate the words that are not providing any precise meaning to the sentence. For example, in fig 1,"They said to concentrate more on practical instead of theory". In this example, the words like "to", "on", "of" will be removed and the remaining words will be segmented. Tweet segmentation helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications. Based on the Parts of Speech (POS) taggers, the tweets are clustered and put under different class labels. Term Frequency is used for the class formation of both global and local context. The classification technique is used for identifying the optimal cluster from different class labels which preserves the actual semantic of the tweets. The rest of the paper is described as follows:- Section 2 represents the literature survey. While section 3 represents the proposed system. Finally, section 4 concludes this paper.

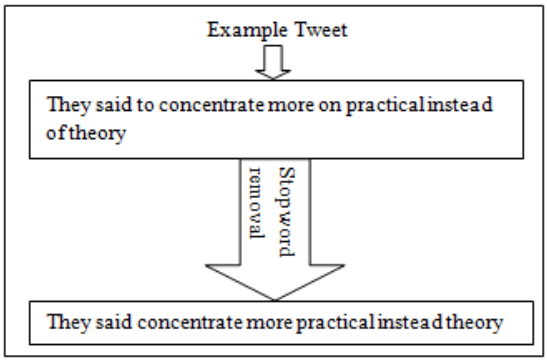

Figure 1: Example for Tweet Segmentation

Some of the applications of the twitter have been described below:-

- **Education:**

Tweets can be useful to provide upcoming due dates for assignments, conducting quizzes & tests, connect with the community (Local Government, charitable organizations) to reach with the broader audience to discuss the latest cultural or educational events, etc.

- **Twitter for Public Relation:**

For Public Relation (PR) professionals and entrepreneurs, there are 3 main reasons to use twitter: Announcements, Research and Networking.

- **Sales & Marketing:**

Twitter is useful in promoting the company brand by optimizing its twitter profile. It is also capable of generating white papers, instructional videos, e-books and product samples if it is suitable.

- **Entertainment:**

People can come closer with many celebrities by following their tweets and images that has been posted by them. This also allows the people to have a talk with their friends and other persons also.

- **Industry & Business:**

Business people use twitter to build their company's brand awareness and thought leadership, to join relevant conversations and talk to like-minded people within their industry and to know what their competitors are up to & to monitor their posts and campaigns.

- **Medical Fields:**

Twitter is rising as a progressively useful tool for Telemedicine and e-health which allows many doctors to keep in touch with patients and fellow professionals.

II. LITERATURE REVIEW

A. Combined analysis of Named-Entity Recognition and Entity Linking

A.Sil [1] analyzed that MSNBC dataset can be used for joining Named- Entity Recognition (NER) system and Entity Linking (EL) system together to make the joint predictions where Named Entity Recognition (NER) is used to find the names which is present in the text, and that names can be connected as the entries in structured or semi-structured repositories like Wikipedia. Entity Linking (EL) system is the process of finding whether a name that appears in text indicates an entity, which appears to be known in an already recognized set of named entities, such as a relational database or the set of articles in Wikipedia. Named Entity Recognition (NER) system cannot able to connect to the Entity Linking (EL) system directly, when NER system failed to detect any mentions. This paper clearly shows that, first NER system will be done and it is followed by the Entity Linking (EL) system. MSNBC dataset has been used for analyzing the results. On using NER it reduces 60% of error and on using EL 68% of error can be reduced.

B. Emoticon Smoothed Language Models for Twitter Sentiment Analysis

The author, K.L.Liu [5] mainly focuses on machine learning based text classification problem which is to mainly identify the attitude or opinion of the tweets. The process of identifying the opinions ("what others think") is called as Sentiment Analysis (SA). For this, [5] some may use manually labeled data to train fully supervised models, while others use some noisy labels, such as emoticons and hashtags, for model training. But in this paper, the author had found that combining both the manually labeled data and noisy labeled data is the best strategy for this approach. They used manually labeled data for providing the training to the language model. This trained model may contain noisy labels which can then be smoothened by using different emoticons. Emoticons can either contain a positive ":)" or negative emoticon ":(" where negative emoticon is not considered. This different emoticon uses some hashtags (eg. #buy) or smileys to identify the sentiment types. However the accuracy of these methods is not satisfactory due to the noisy labels.

C. Opinion Summarization using hashtags and Human annotated semantic tags in Twitter

The author, X.Meng [6] extracted how the automatic opinion summarization poses a greatest challenge to the summarization system by considering the users opinions or attitude which are posted as the tweets in the twitter. The

main focus of the paper is to identify the specific topic related opinion summaries such as [6] celebrities and brands. They used hashtags and human annotated semantic tags for calculating the similarity among them to provide better interpretation and representation. Next, they grouped #hashtags into coherent topics by adopting affinity algorithm. Then they focused on the entity related opinions. ie; when an entity is provided they collected the opinions for it. Finally, the summary is generated from different opinions which can be based on various topics and opinions produced by the users in the twitter.

D. Automatic detection of Named Entities using Dynamic Programming & Random Walk model

C.Li [3], analyzed that the author has concentrated on Targeted Twitter Stream which requires a Named Entity Recognition (NER) system to discover the named entities automatically. This Targeted Twitter Stream is used for monitoring, understanding and collecting the user opinions about the organizations. This paper presented a TwiNER concept where NER system has been used and it comes under unsupervised learning algorithm. First, it takes global context from Wikipedia and uses dynamic programming algorithm to divide the tweets into valid segments where each of these segments are called as candidate named entity. Secondly, Random walk model is constructed by the TwiNER, to avoid the gregarious property of local context. Then it finally ranks each segments and the segment with highest ranking will be considered as a true named entities.

E. Structuring Topical N-grams model for topic and its topical phrases

The author, X.Wang [7] concentrates on the situations where the meaning of the text is difficult to understand due to the order of words and phrases present in the text. This is called as bags of words assumption. The paper provides topical n-grams, which is modeled to discover [7] topics as well as topical phrases. Each word in the textual ordered sentence is sampled at first and it identifies the status of a word by checking whether it is a unigram or a bigram. This paper mainly focuses on determining the unigram words and phrases automatically by constructing Topical-n Gram (TNG) model. This model is also developed to provide meaningful phrases by helping linguists.

F. Lexical Normalisation of Short Text Messages

B.Han [2] described vocabulary words which are provided as a short text messages that are called as ill-formed words. To overcome the difficulty of understanding those words, the author developed a classifier to detect such ill-formed words and normalize it based on [2] morphophonemic similarity. The aim of this paper is the task of lexical normalisation of noisy English text, with a particular focus on Twitter and SMS messages. OOV word distribution of Twitter and other text genres is to be studied for providing a [2] text normalization dataset based on twitter data. Finally, it makes a word similarity check, dictionary lookup to make the word understandable quickly.

Table I. Table Comparison For Different Techniques

Methods	Functionalities	Advantages	Disadvantages
Joint Named-Entity Recognition and Linking	Named Entity Recognition (NER) system and Entity Linking (EL) system are joint together to make the predictions on different mentions present in the text.	System is computationally inexpensive in the predicting and linking the recognized entities.	Collective classification Provide only small benefits on the purely local modal.
Emoticon Smoothed Language Models	To understand the different sentiment types (i.e.; identifying various opinions or attitude of the tweets), the paper uses both manually labeled data and noisy data.	Classification of the labeled data for sentimental analysis yields high precision and recall value.	Noisy labeled data has not pruned for classification with high accuracy.
Hashtags and human annotated semantic tags	The paper focuses on collecting the opinions of specific topic and it generates a summary automatically. It uses hashtags and human annotated semantic tags.	Identification of perceptive tweets with high precision is possible. Target dependent sentiment classification is used to identify different opinions.	It is impossible to extract the complete and the most interesting topics tweeted. Each opinion must be analyzed every time and therefore, it takes more time for its execution.
TwiNER: Named Entity Recognition	To monitor, understand and collect different user opinions about the organization by using	User defined selection a criterion is applied for filtering tweets on local context with high	Entity type classification is not addressed.

	Targeted Twitter Stream. It uses 2-step unsupervised NER system called TwiNER.	scalability.	
Topical N Gram (TNG) model	To overcome the difficulty of understanding the meaning of disordered words and phrases in a text the authors constructed, Topical N-gram Model (TNG) model. It identifies whether the word in the text is a unigram or bigram.	Bigram method is used for longer sentences for decomposing and classification process.	It uses a statistical pruning logic which leads to failure results in classification.
Lexical Normalisation using morphophonemic similarity	To detect the ill-formed words, classifier is used and it can be normalized using morphophonemic similarity.	It uses word similarity and word context to detect ill-formed words which is easy for the users to understand	It doesn't alleviate the noisy contents. There is a possibility where text normalization may failed to detect some ill-formed words

III. PROPOSED SYSTEM

Initially as in fig.2, tweet data taken from the twitter data source will be preprocessed by dividing the tweets into valid segments, to remove the word which does not have a specific meaning to be provided to understand that word. This can be done using stop word removal and stemming process. Then the preprocessed data will be parsed for providing the POS tagging and put under different class labels. Afterwards, the tweet Classification can be obtained from feature selection. Tweet data taken and parsed must be checked for identifying whether it is a local or global context, to determine the feature selection. Finally, the number of times a word occurs in each class label can be calculated by using term frequency and linguistic features. Term frequency generates the offset value for the same words that are repeated frequently. This offset value by the frequency of the word increases proportionally for each occurrences of the same word. Linguistic features uses NLP tool to map an input text with counts on dictionary-supplied categories.

Apart from the previous mentioned techniques some of the other techniques involved in the paper are described below:-

A. Global Context

Tweets containing normal English words obtained from Wikipedia that can form a meaningful sentence which is called as a Global Context. This tweets can be partitioned into various segments which is therefore useful for preserving the meaning of each segments present in the tweets. Named Entities and semantic phrases can be easily identified in the tweets.

B. Local Context

Tweets containing any short text, ill-formed words or grammatical mistakes are said to be called as a Local Context. There are many tweets posted every minute in a day at a faster rate therefore, this local context may not be understandable to many users those who are not well known with the short text messages. It does not provide a meaningful phrase even the tweets are partitioned into valid segments.

Example:"He playin"

C. Linguistic Features

Linguistic features such as expressivity, complexity, affect, informality, uncertainity, non-immediacy, diversity and emotional consistency have extraordinary potential for human deception detection in text based communication. NLP is a tool which is used to detect the linguistic features. It is a tool which maps an input text with counts on dictionary supplied categories.

D. Parts of Speech(POS) tagging

POS is the task of identifying the category to which a word is assigned in accordance with its syntactic functions. A data must be tagged using POS tagger after the parsing mechanism.

E. Named Entity Recognition(NER)

Named Entity Recognition (NER) labels are sequences of words in a text which can be a names of the persons and company names. It is the subtask of information extraction that seeks to locate and classify the elements in text into predefined categories such as the names of the persons, organizations etc. NER is also called as entity identification (or) entity extraction.

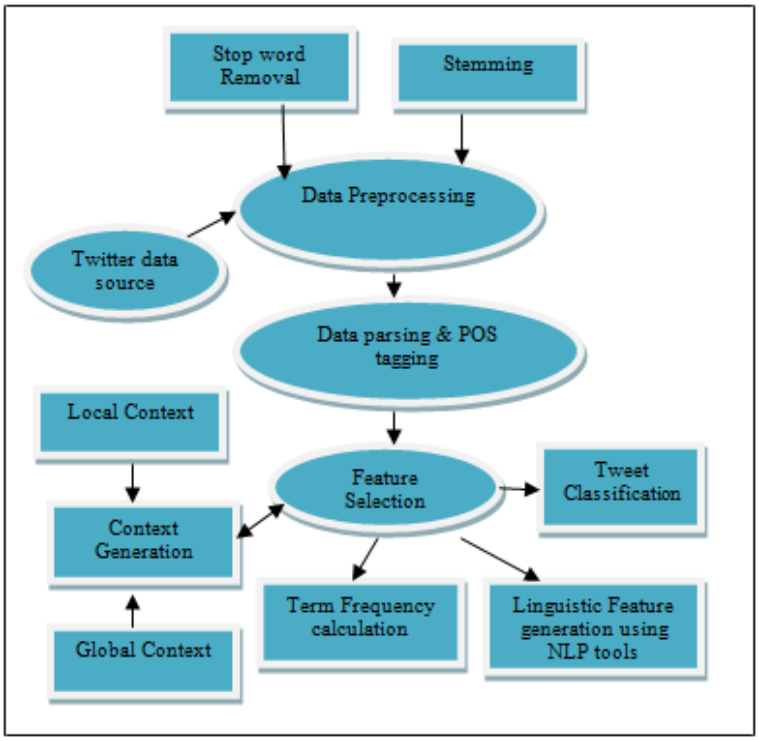

Figure 2: Architecture Diagram

F. Term Frequency

Term Frequency is a numerical statistic data which identifies the number of occurrences of each word in the document by creating the offset value. This shows the importance of that particular word to the document. It is used as the weighting factor in retrieving the information and text mining.

IV. CONCLUSION

The main aim of this paper is to provide the opinion summary which can reduce the execution time and make this method more reliable. This present study deals with how the tweet summarization can be made to avoid the overload problems due to the diversities among the sentences. The opinion summarization can be generated by considering both the global and local context. It is done by using different methods like data preprocessing, Parts of Speech (POS) tagging and tweet classification. This can be helpful for the users to understand the tweets easily.

REFERENCES

[1] A. Sil and A. Yates, "Re-ranking for joint named-entity recognition and linking," in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.,2013, pp. 2369–2374

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

[3] C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee, "Twiner: Named entity recognition in targeted twitter stream," in Proc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012, pp. 721–730.

[4] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, Member, "Tweet Segmentation and Its Application to Named Entity Recognition", IEEE transaction, VOL. 27, NO. 2, FEBRUARY 2015.

[5] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in Proc. AAAI Conf. Artif.Intell., 2012, pp. 1678–1684.

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

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

[8] Zhenhua Wang, Lidan Shou, Ke Chen, Gang Chen, and Sharad Mehrotra, "On Summarization and Timeline Generation for Evolutionary Tweet Streams", IEEE transaction, VOL. 27, NO. 5, MAY 2015.

