Tweet data Preprocessing and Segmentation to NER

Mr.Sanket Patil, Prof. Varsha Wangikar, Prof. K. Jayamalini

Abstract— Social media offers a powerful outlet for people's thoughts and feelings. Now-a-days social networking sites are at the boom, so large amount of data is generated. Millions of people share their views daily on micro blogging sites, since it contains short and simple expressions. In this project, we will discuss about a paradigm to extract the data from a famous micro blogging service, Twitter, where users post their opinions for everything and NER will be applied on the same to classify different name entities. Here, we will discuss an approach that automatically classifies the sentiments of Tweets taken from Twitter dataset. Tweets are classified as positive, negative or neutral with respect to a query term. Tweet data taken from the twitter data source will be pre-processed 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 helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications. Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. We show that high accuracy in named entity recognition is achieved by applying the method Stanford NER. Together, this is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product along with the extraction of geographical location from tweets.

Index Terms—*Empirical Distribution, Microblogging, NER, Segmentation ,Sentiment Analysis, ,Social media, Twitter.*

*****

I. INTRODUCTION

In today's world, virtual communities and networks on Internet are adopted so well that even a term is derived for them i.e. Social Media. Social media forms such as social networking or micro blogging websites allow people to create and share any kind of information and ideas. Facebook, twitter etc. are widely used forms of social media. Twitter, has already settled into our lives and has become one of the most important communication channels with its ability of providing the most up-to-date and newsworthy information. To collect and understand user's opinion many organizations have been reported to create and monitor twitter streams. Targeted data in twitter stream is usually build by filtering tweets with predefined selection criteria (e.g., tweets published by users from a particular geographical region, tweets that matches one or more predefined key- words).It is imperative to understand tweets language for a large body of downstream applications, such as named entity recognition (NER),opinion mining, sentiment analysis and many others due to its invaluable business value of timely information from these tweets, Tweet is given a limited length i.e. 140 characters and no restrictions on its writing styles which make tweets contain with grammatical errors, misspellings and informal abbreviations. This short nature and error prone tweets often make the word-level language models for tweets less reliable. For example, given a tweet "I call him, no answer. His phone in the bag, he dancing," there is no clue for guessing its true theme by ignoring word order (i.e., bag-of-word model). Further the situation is made worse with limited context provided by the tweet. That is, if the tweet is considered in isolation than more explanation or conclusion can be deduced. While on the other hand, with noisy nature of the tweets the core semantic information is well preserved in form of named entities or semantic phrases in the given tweets. Positive, negative, neutral are the categories into which the area of Sentiment Analysis intends to comprehend these opinions and distribute them Till now most sentiment analysis work has been done on review sites [. Review sites provide with the sentiments of products or movies, thus, restricting the domain of application to solely business. Sentiment analysis on Twitter posts is the next step in the field of sentiment analysis, as tweets give us a richer and more varied resource of opinions and sentiments that can be about anything from the latest phone they bought, movie they watched, political issues, religious views or the individuals state of mind. Thus, the foray into Twitter as the corpus allows us to move into different dimensions and diverse applications.

II. REVIEW OF LITERATURE

S. Vijayarani [1] have published the research on text mining and its preprocessing techniques. Text mining is the process of mining the useful information from the text documents. It is also called knowledge discovery in text (KDT) or knowledge of intelligent text analysis. Text mining is a technique which extracts information from both structured and unstructured data and also finding patterns. Text mining techniques are used in various types of research domains like natural language processing, information retrieval, text classification and text clustering. Text mining is the process of seeking or extracting the useful information from the textual data. It tries to find interesting patterns from large databases. It uses different pre-processing techniques likes stop words elimination and stemming.

Swati B. Bhonde [2] has proposed a novel algorithm for the preprocessing of the text using distributed approach is introduced. This algorithm makes use of many techniques to remove redundant information used for sentiment analysis and it will result in compressing a lengthy statement into shorter and easier one without changing its meaning. Filtered sentences enhance the performance of sentiment analysis system in terms of time & space complexity also. The domain specific application gives more correct results for text mining. This paper presented pre-processing techniques using distributed approach, which significantly reduces time required in preprocessing phase.

Chenliang Li [3] proposed a novel framework for tweet segmentation in a batch mode, called HybridSeg is proposed. By splitting

tweets into meaningful segments, the semantic or context information is well preserved and easily extracted by the downstream applications. HybridSeg finds the optimal segmentation of a tweet by maximizing the sum of the stickiness scores of its candidate segments. The stickiness score considers the probability of a segment being a phrase in English (i.e., global context) and the probability of a segment being a phrase within the batch of tweets (i.e., local context). For the latter, two models are proposed and evaluated to derive local context by considering the linguistic features and term-dependency in a batch of tweets, respectively. HybridSeg is also designed to iteratively learn from confident segments as pseudo feedback. Experiments on two tweet data sets show that tweet segmentation quality is significantly improved by learning both global and local contexts compared with using global context alone. Through analysis and comparison, it is shown that local linguistic features are more reliable for learning local context compared with term-dependency. As an application, it is shown that high accuracy is achieved in named entity recognition by applying segment-based part-of-speech (POS) tagging.

Kamal Nigam proposes the use of maximum entropy techniques for text classification. Maximum entropy [4] is a probability distribution estimation technique widely used for a variety of natural language tasks, such as language modeling, part-of-speech tagging, and text segmentation. The underlying principle of maximum entropy is that without external knowledge, one should prefer distributions that are uniform. Constraints on the distribution, derived from labeled training data, inform the technique where to be minimally non-uniform. The maximum entropy formulation has a unique solution which can be found by the improved iterative scaling algorithm. In this paper, maximum entropy is used for text classification by estimating the conditional distribution of the class variable given the document. In experiments on several text datasets accuracy to naive Bayes is compared and show that maximum entropy is sometimes significantly better, but also sometimes worse. Much future work remains, but the results indicate that maximum entropy is a promising technique for text classification.

In [5] A. Agarwal et.al. proposed a paradigm to extract the sentiment from a famous micro blogging service, Twitter, is discussed, where users post their opinions for everything. Further discussed is the existing analysis of twitter dataset with data mining approach such as use of Sentiment analysis algorithm using machine learning algorithms. An approach is introduced that automatically classifies the sentiments of Tweets taken from Twitter dataset. These messages or tweets are classified as positive, negative or neutral with respect to a query term. This is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product before purchase. Machine learning algorithms are used for classifying the sentiment of Twitter messages using distant supervision. The training data consists of Twitter messages with emoticons, acronyms which are used as noisy labels. Sentiment analysis on Twitter data was examined. The contributions of this survey paper are: (1) Use of Parts Of Speech (POS)-specific prior polarity features (2) Use of a tree kernel to prevent the need for monotonous feature engineering.

Shilpy Singh [6] published the research on the utilization of Twitter4J libraries is explored which are reliable Twitter APIs and that can be integrated to any applications for data acquisition in any format. It is a cross-platform tool and can be used on several operating systems, with the latest versions of Java Runtime Environment. The utility can be used as it is without any customizations it has no dependencies to any other system on which it runs. The usage of Twitter4J is simple, as all one needs to do is copy the JAR file to the preferred class path and use it. The method of using twitter4J libraries for data acquisition for data analytics is explored here. This work will help data scientist, data quality analyst and business users. The usage of machine learning techniques to perform for classifying sentiment in tweets is emphasized here. Also demonstrated how one can use twitter4j for streaming. It also supports programmatic access to the actions that any Twitter user can take, including posting messages, retweeting, following, and more. In this work, we have demonstrated a system to extract knowledge from tweets and then classify tweets based on the semantics of knowledge contained in them. For avoiding information loss, knowledge enhancer is applied that enhances the knowledge extraction process from the collected tweets. The maturity of knowledge gained using knowledge enhancer module has helped to filter tweet more precisely avoiding information loss. Also measured missing information during specific keyword-based search and then proposed a method to collect more precise information about specific topic or domain. Sentiment analysis shows people attitude towards different topics. This data can also help to generate richer user profile and generate valuable recommendations.

In [7] author proposed to use the Twitter corpus to ascertain the opinion about entities that matter and enable consumption of these opinions in a user friendly way. The focus is on classifying the opinions as positive, negative or neutral. Since there aren't large enough datasets of labeled tweets, limiting the sentiment categories to the above three enables us to leverage other similar but larger datasets for training custom sentiment language models. It was begun by extracting entities from the Twitter dataset using the Stanford NER.

2.1 Data Characteristics

The Twitter is a social networking and microblogging service that lets its users post real time messages, called tweets, Implication of new challenges and shape up the means of carrying sentiment analysis on it as compared to other domains is due to tweets having unique characteristics.

Following are some key characteristics of tweets:

1. *Message Length:* The maximum length of a Twitter message is 140 characters. This is different from previous sentiment classification research that focused on classifying longer texts, such as product and movie reviews.
2. *Writing technique:* In comparison with other domains the occurrence of incorrect spellings and cyber slang in tweets is more often . As the messages are quick and short, people use acronyms, misspell, and use emoticons and other characters that convey special meanings.
3. *Availability:* The amount of data available is immense.Data is more readily available as more people tweet in the public domain as compared to Facebook (as Facebook has many privacy settings) . The Twitter API facilitates collection of tweets for training.
4. *Topics:* Twitter users post messages about a range of topics unlike other sites which are designed for a specific topic. This differs from a large fraction of past research, which focused on specific domains such as movie reviews.
5. *Real time:* Blogs are updated at longer intervals of time as blogs characteristically are longer in nature and writing them takes time. Tweets on the other hand being limited to 140 letters and are

updated very often. This gives a more real time feel and represents the first reactions to events.

We now describe some basic terminology related to twitter:

1. *Emoticons:* These are pictorial representations of facial expressions using punctuation and letters. The purpose of emoticons is to express the user's mood.
2. *Target:* Twitter users make use of the "@" symbol to refer to other users on Twitter. Users are automatically alerted if they have been mentioned in this fashion.
3. *Hash tags:* Users use hash tags "#" to mark topics. It is used by Twitter users to make their tweets visible to a greater audience.
4. *Special symbols*: "RT" is used to indicate that it is a repeat of someone else's earlier tweet.

III. SYSTEM ARCHITECTURE

A Opinion words are the words that people use to express their opinion (positive, negative or neutral). To find the semantic orientation of the opinion words in tweets, we propose a novel approach involving dictionary-based techniques. We also consider features like emoticons and capitalization as they have recently become a large part of the cyber language. Standford NER is applied on twitter data set to classify the different entities.

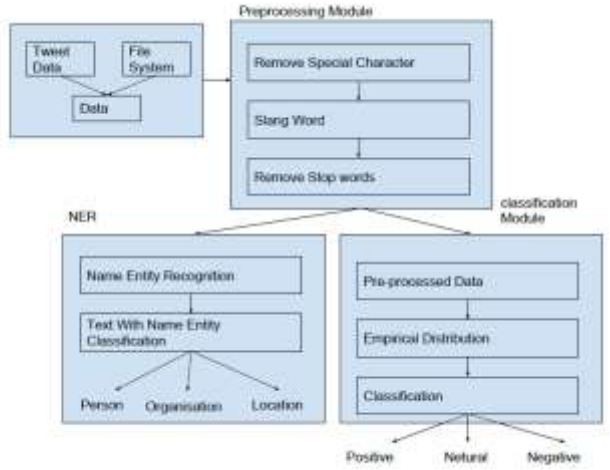

Fig. 1 Proposed Architecture of tweet data pre-processing and segmentation to NER.

1. Data gathering:

Data gathering is the first part in which we provide the data for further processing in the application. Data can be provided from online as well as offline mode. Online data is downloaded from the Twitter website by creating Twitter API and passing account credential. input which gets stored on the file system for further processing. Offline data is the data which is saved on the disk in a text file. This file's path is given as input to the application which is then extracted by the application and stored on file system.

2. Data pre-processing:

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. In this application data pre-processing involves three step which includes the following:

a) Data cleaning: Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied. Therefore, a useful pre-processing step is to run your data through some data cleaning routines which includes filling up with missing values and reducing noisy data.

b) Removing Slang words:

One of the first steps in working with text data is to pre-process it. It is an essential step before the data is ready for analysis. Majority of available text data is highly noisy and unstructured in nature – to achieve better insights or to build better algorithms, it is necessary to play with clean data. For example, data is highly unstructured in social media – it is an informal communication – typos , usage of slang,bad grammar, presence of unwanted content like URLs, stop words, Expressions etc. are the usual suspects. Internet slang consists of slang and acronyms that users have created as an effort to save keystrokes.

c) Removing stop words:

Stop words are basically a set of commonly used words in any language, not just English. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead. Stop words are generally thought to be a "single set of words". It really can mean different things to different applications. For instance, in sentiment analysis removing adjective terms such as 'good' and 'nice' as well as negations such as 'not' can throw algorithms off their tracks. In such cases, one can choose to use a minimal stop list consisting of just determiners or determiners with prepositions or just coordinating conjunctions depending on the needs of the application.

Examples of minimal stop word lists that you can use:

- Determiners - Determiners tend to mark nouns where a determiner usually will be followed by a noun examples: the, a, an, another
- Coordinating conjunctions – Coordinating conjunctions connect words, phrases, and clauses examples: for, an, nor, but, or, yet, so
- Prepositions - Prepositions express temporal or spatial relations examples: in, under, towards, before

In some domain specific cases, such as clinical texts, we may want a whole different set of stop words. For example, terms like "mcg" "dr" and "patient" may have less discriminating power in building intelligent applications compared to terms such as 'heart' 'failure' and 'diabetes'. In such cases, we can also construct domain specific stop words as opposed to using a published stop word list.

3. Classification Module:

For classification two methods are applied as follows:

a) Every tweet is checked for the positive and negative words from a fixed set of list. Then the average is taken out for both positive and negative, depending on the higher score the tweet label is saved as positive /negative or if the score is 0, it's neutral.
b) The tweet words are also checked from sentiword.net and the score is again calculated for every tweet and labeled as positive if greater than 0 and labeled as negative if less than 0.

Then the two labels are compared if both the tweets share the same sentiment the label does not change, but if the labels are opposite then the score from each method is checked and the higher score label is assigned.

For eg, "I love India but India is not that clean"

Method 1: positive Method 2: negative

Method1: score= 0.9 Method2: score= -1.9

|method1:score|<|method2:score|

Tweet is labeled as negative.

4. Name entity recognition

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. We show that high accuracy in named entity recognition is achieved by applying the method Stanford NER. Together, this is very useful for the companies who want to know the feedback about their product brands or the customers who want to search the opinion from others about product along with the extraction of geographical location from tweets.

Stanford NER is a Java implementation of a Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names.Stanford NER is also known as CRFClassifier. The Stanford NER can't automatically classify names tags (@person) and URLs as entities so we augment it by including these types of entities. It seems intuitive to collect person tags since this is how Twitter users express their opinion or communicate with the entity being tagged. Similarly, opinions about URLs that are shared on Twitter, since disseminating URLS is of the chief use cases for Twitter.

IV. RESULTS & DISCUSSIONS

We applied our approach to a sample set of tweets. The semantic analysis results obtained are depicted in Table 1 below:

TABLE 1
SEMANTIC ANALYSIS ON SAMPLE TWEETS

Tweet	Orientation	Score
hands amir khan champion movie maker present times pick winning team play game	Positive	1.001
youre feeling sad remember amir khan called daughter process	Negative	-0.82
billgates Microsoft expects tough market for Windows 7: Microsoft executive Bill Koefoed answered some questions about the..	Negative	-0.812
i hate bollywood movies amir khan exceptional case	Negative	-0.779
amir khan fight saul alvarez world title las vegas	Negative	-0.82
dangal movie review critics	Neutral	0

Tweets and NER:

Example 1:

@billgates Microsoft expects tough market for Windows 7: Microsoft executive Bill Koefoed answered some questions about the.. http://tinyurl.com/l4a4z9

NER:

<NAME>@billgates</NAME><ORGANIZATION>Microsoft</ORGANIZATION>expects tough market for Windows 7: <ORGANIZATION>Microsoft</ORGANIZATION> executive <PERSON>Bill Koefoed</PERSON>answered some questionsabout the..<URL>http://tinyurl.com/l4a4z9</URL>

Example 2:

Amir khan fight saul alvarez world title las vegas

NER:

<name>amir </name><name>Khan </name> fight <name>saul </name><name>alvarez </name> world title<Location>lasvegas</location>

V. CONCLUSION

Microblogging sites like Twitter offers an opportunity to create and employ theories & technologies that search and mine for sentiments. The work proposed in this paper specifies a novel approach for sentiment analysis on twitter data. When users post their opinions, Standford NER is applied on the twitter data to classify the different name entities. Dictionary based method and maximum entropy technique was used to find the semantic orientation of verbs and adverbs which improved the accuracy of the classification.

REFERENCES

[1] S. Vijayarani,J. Ilamathi,Nithya, " Preprocessing Techniques for Text Mining - An Overview" *International Journal of Computer Science & Communication Networks,Vol 5(1),7-16 ISSN:2249-5789*

[2] Swati B. Bhonde, Jayashree R. Prasad *"Improved Approach for Preprocessing Text using Hadoop framework*, ISBN : 978-0-9948937-4-1

[3] Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, "Tweet Segmentation and Its Application to Named Entity Recognition", IEEE Transactions on Knowledge and Data Engineering, 2015.

[4] Kamal Nigam, John La_erty, Andrew McCallum "Using Maximum Entropy for Text Classi_cation".

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau,"Sentiment Analysis of Twitter Data", In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 ,pp. 30–38S.

[6] Shilpy Singh,Manjunath T N, Aswini N," A Study on Twitter 4j Libraries for Data Acquisition from Tweets" *International Journal of Computer Applications (0975 – 8887) National Conference on "Recent Trends in Information Technology" (NCRTIT-2016*

[7] Batra and D. Rao, "Entity Based Sentiment Analysis on Twitter", Stanford University,2010