A EXPLORATORY ANALYSIS OF PROCESSING FREQUENCY WORD DENSITY ALONG WITH NAMED ENTITY RECOGONITION

S.Mythrei*[1], S.Singaravelan[2] , R.Arun[3] , D.Murugan[4]
*1Research Scholar, Department of Computer Science and Engineering, PSR Engineering College, Sivakasi, India.
2,3 Associate Professor, Department of Computer Science and Engineering, PSR Engineering College, Sivakasi, India.
4 Professor, Department of Computer Science and Engineering, Manonmaniam Sundaranar University ,Tirunelveli, India.
mythu.sri91@gmail.com , singaravelan.msu@gmail.com

ABSTRACT

Text analytics is the most popular form in our day to day conversion. Most of the data which we generate is unstructured and leads into processing the generate insights in nature. Natural language processing enables the human to interact with systems in a natural way. Named Entity Recognition (NER) is task of extracting information from unstructured text which can be categorized by persons, locations, organizations, cost values, percentages, expressions and so forth. This paper describes the processing of the twitter positive sentiment tweets and negative sentiment tweets, text corpus. The tweets and web text corpus data noises are pre processed, frequency density of words are analyzed from tweets and web text corpus data by following with Named Entity Recognition(NER) chunked tree for twitter corpus and web text corpus.

Keywords Frequency Density, Tweets, Sentiments, Named Entity Recognition

1. INTRODUCTION

The amount of data has been increased exponentially and the web becomes one of the largest repositories for data. The web hold more amount of web data forms the natural languages. Unstructured data is essentially important which has a internal structure via pre-defined schema or data models. Unstructured data can be of images, videos, text corpus and audio. Natural language Processing (NLP), text mining, pattern classification and pattern sensing are most common examples for corpus analytics, sentiment analytics, finding entities from the corpus sentences.NER can be used in so many fields in NLP which can help in answering real-world questions like which organisation were mentioned in this article, Does this twitter corpus contain the persons name, location, organization and so forth.

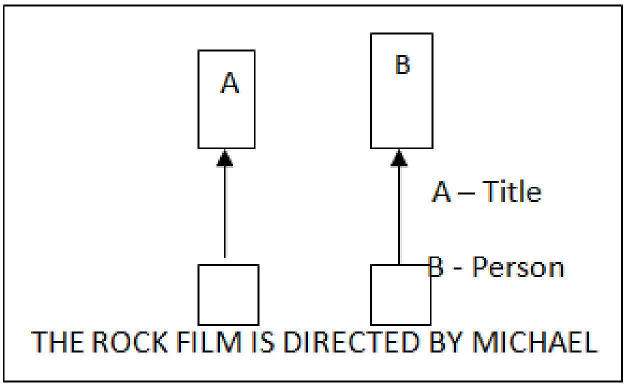

Fig.1. Example for NER

In the Fig.1, the words 'The Rock' and 'Michael' represent the NER. The tweets and text corpus contan many informal abbreviations and grammatical mistakes as it is an unstructured data that there is no constraints in its writing style.

2. RELATED WORK

.In [1] proposed a novel multi-level architecture that does not rely on any specific linguistic resource or encoded rule and to use features extracted from images and text to classify named entities. a multi-level architecture which intends to produce biased indicators to a certain class (LOC, PER or ORG). These outcomes are then used as input features for final classifier. It's a novel architecture for NER that expands the feature set space based on feature clustering of images and texts, focused on micro blogs. Due to their terse nature, such noisy data often lack enough contexts, which pose a challenge to the correct identification of named entities. To address this issue have been evaluated a novel approach using the Ritter dataset. In [2] describes the range of different features were developed to extract Twitter names from the tweets. Two systems were built, one for the 'no types' named entity extraction task and the other for the '10types' classification task. The systems were built around a CRF-based classifier and lexical data, and both systems achieved state-of-the-art results. Twitter named entity recognition is the process of identifying proper names and classifying the min to some predefined labels/categories. The paper introduces a Twitter named entity system using a supervised machine learning approach, namely Conditional Random Fields. A large set of different features was developed and the system was trained using these. . In [3] presented a novel 2-step unsupervised NER system for targeted Twitter stream, called TwiNER. In the first step, it leverages on the global context obtained from Wikipedia and Web N-Gram corpus to partition tweets into valid segments (phrases) using a dynamic programming algorithm. Each such tweet segment is a candidate named entity. The highly-ranked segments have a higher chance of being true named entities. We evaluated TwiNER on two sets of real-life tweets simulating two targeted streams. A region-based stream for tweets published by users from a particular geographical region; and a topic-based stream for tweets potentially relevant for a political event.In [4] explained each entity string in our data is associated with a bag of words found within a context window around all of its mentions, and also within the entity itself. A plethora of distinctive named entity types are present, necessitating large amounts of training data. T-NER system doubles F1score compared with the Stanford NER system. T-NER leverages the redundancy inherent in tweets to achieve this performance, using Labelled LDA to exploit Freebase dictionaries as a source of distant supervision

3. EXPERIMENTS

3.1. DATASET

NLTK is a python package which holds a set of diverse NLP algorithms. It is open source, free, ease of use and well documented. We choose twitter samples and webtext corpus from NLTK dataset. The tweets dataset is a part of NLTK package that has been downloaded and imported for our work. It contains set of tweets for positive and negative sentiment tweets and tweets. The web text dataset is also a part of NLTK package that contains six text corpus where we use two text corpus data in our work.

3.2. **EFFICIENCY**

In this subsection, we have evaluated the frequent occurrence of the words in the positive sentiment twitter dataset, negative sentiment twitter dataset, frequent occurrence of letters in text corpus from web text dataset. Along with this, the NER for positive sentiment twitter dataset and negative sentiment twitter dataset can be analyzed. A set of 5000 positive and negative sentiments tweets and 20000 tweets along with 65003 text grail text corpus used to train and test our model for finding frequent occurrence of words. Initially the data are pre-processed by tokenized the sentiment tweets and corpus text, the words have been stemmed and lemmatized. The tweets data are unstructured in nature which contains noise and the common words in English language are stop words.

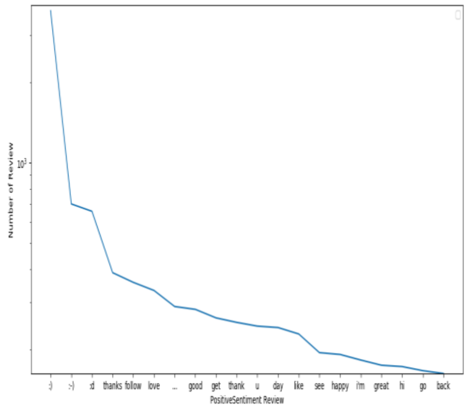

Fig.2.Frequency of words in Positive sentiment tweets

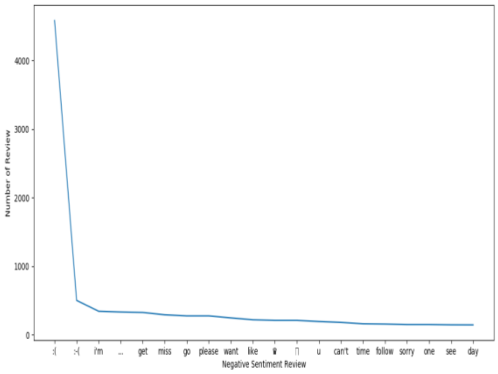

Fig.3.Frequency of words in Negative sentiment tweets

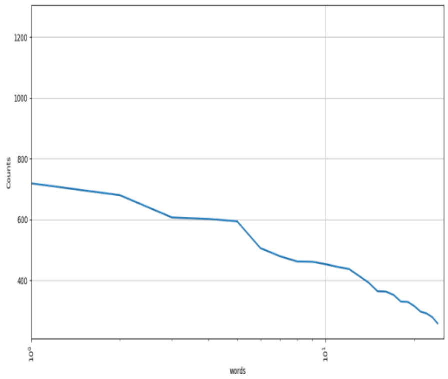

Fig.4.Frequency of words in Web text corpora

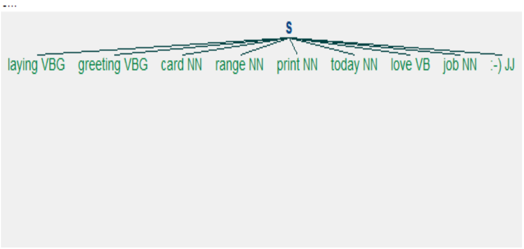

Fig.5. NER chunked graph for Positive sentiment sentiment Tweets

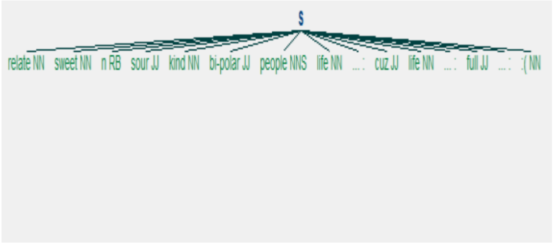

Fig.6. NER chunked graph for Negative Tweets

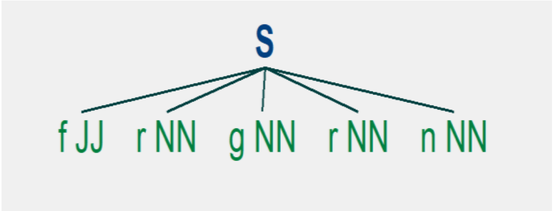

Fig.7. NER chunked graph for Grail text corpus

In our dataset, the hyperlinks, special characters and punctuations are considered as noise which can be removed by using regular expression library. Fig.2 shows the frequency of words analysis over positive and negative sentiment tweets. Fig.3 shows the frequency of words analysis over negative sentiment tweets. From Fig 4, we can observe the frequent occurrence of words from web text corpus.

The named entities in the positive and negative sentiment tweets are find out from the chunks of the data. The chunked method in NLTK generates a tree structure for the named entities. Whereas the String S as the root node and the child note of the tree represents the named entity nodes of sentiment tweets. Fig.5.represents the chunked graph for Positive sentiment tweets.Fig.6.represents the NER chunked graph for Negative sentiment tweets. Fig.7. represents the NER chunked graph for Grail text corpus

4. CONCLUSION

In this paper, we have a studied a problem of pre processing the tweets and web text corpus how to remove noise from the same. The experimental result shows that the analysis of frequency word density for positive sentiment tweets, negative sentiment tweets and web text corpus data. Through the extensive experiments, Named Entity Recognition (NER) chunked tree structure is constructed for twitter sentiments and grail web text corpus.

5. REFERENCES

1. Diego Esteves1, Rafael Peres2, Jens Lehmann1,3, and Giulio Napolitano, Named Entity Recognition in Twitter using Images andText, International Conference on Web Engineering ICWE 2017: Current Trends in Web Engineering pp 191-199, February 2018.

2. Utpal Kumar SikdarandBjorn Gamback, Feature-Rich Twitter Named Entity Recognition and Classification, Proceedings of the 2nd Workshop on Noisy User-generated Text,pages 164–170, Osaka, Japan, December 11 2016.

3. Chenliang Li, Jianshu Weng,QiHe,YuxiaYao, Anwitaman Datta,Aixin Sun, and Bu-Sung Lee, TwiNER: Named Entity Recognition in Targeted TwitterStream, SIGIR’12,ACL,2012.

4. Alan Ritter, Sam Clark, Mausam and Oren Etzioni, Named Entity Recognition in Tweets:An Experimental Study, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 1524–1534,JULY 2011.

5. LIU, X., ZHOU, M., WEI, F., FU, Z.,ANDZHOU, X. Joint inference of named entityrecognition and normalization for tweets. InProceedings of the 50th Annual Meeting of theAssociation for Computational Linguistics: Long Papers-Volume 1(2012), Association forComputational Linguistics, pp. 526–535.

6. Derczynski, L., Maynard, D., Rizzo, G., Vanerp, M., Gorrell, G., Troncy, R.,Petrak, J.,Andbontcheva, K. Analysis Of Named Entity Recognition And Linking For Tweets. Information Processing & Management 51, 2 (2015), 32–49.

7. Doug Downey, Matthew Broad head, and Oren Etzioni.2007. Locating complex named entities in web text. In Proceedings of the 20th international joint conference on Artifical intelligence.

8. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In Proc. of ACL, 2002.