International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 2033-2037
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Normalization of Text in Social Media: Analyzing the Need for Pre-processing Techniques and its Roles

Nimala.K
Department of Information Technology, SRM University, India,Email-nimskt@gmail.com.
Dr.Thenmozhi. S
Department of Information Technology, SRM University, India,Email-mozh_2000@yahoo.com.
Dr.ThamizhArasan. R
Department of Computer science, Bharath University, India, Email-eniyan2000@yahoo.com

Abstract —**Nowadays social media have changed our day to day lives. Social media could be stated as a kind of website or an application where people tend to share their thoughts, ideas, views etc. among individuals or group's .It is a kind of electronic communication where people create online communities to share information, ideas, personalmessages, videos etc. Often the social media data is unsuited for analysis due to the irregularity of the languagefeatured. This paper focuses on the role of pre-processing of social media data which would enable better classifying of text for sentiment analysis, topic modelling etc. The paper contributes the various aspects of pre-processing the large massive datasets obtained from social network such as twitter, face book, LinkedIn etc.**

Keywords - Social Media, Sentiment analysis, Machine Learning, classification

I. INTRODUCTION

Social media is a way for people to interact and communicate online. It is calledas social media because users engage with it in a social context which includes commentary, conservations, user generated annotations etc.Very often the massive data generated is noisy and not suited for classification. So there comes an objective of how to improve the utility of social media text. The basicidea that comes into one's mind is the divide and conquer strategy. The technique follows as,divide massive data into parts where parallel computing architecture could be implemented and conquer the noisy data to re-build NLP Pipelines, transform to make it more accessible to the existing tool. Usually online Text contains lots of uninformative parts such as HTML tags, scripts,special symbols, stop words, white spaces,noisesand advertisements; in addition to it the words used in the online text do not have impact on the general orientation of it. So the problem under this context is high and hence the classification is more difficult on it. In order to have real sentiment analysis, the role of pre- processing the text gets its importance.

Pre-Processing is the process of cleaning or in other words scrubbing and preparing the data for effective analysis of data. By cleaning up the data, the noise level in the text reduces, thereby it speedups classification and improves the performance of the classifier. Section II discusses on the related works carried out in pre-processing of short text and section III focuses on the various techniques/methodology involved in tweet processing and section IV talks about the experimental setups and results and further Section V speaks on the conclusion and future works.

II. RELATED WORKS

With dramatic increase of users on social media, their rises massive data being generated. So this massive data could be made useful by extracting useful knowledge out of it. The preliminary challenge lies in pre-processing the large data. Various Techniques have being followed by analyst in processing such massive data. S. Vijayarani [1] provided the overview of text mining and its importance and need of preprocessing techniques.

Giulio Angiani provides the comparison of different preprocessing techniques on the dataset taken from twitter.[2] conclude that using dictionary did not enhance the performance, rather it increased the elaboration time needed for cleaning raw data. Emma Haddi et al explored the role of text preprocessing in sentiment analysis and experimented that feature selection and representation along with SVM classification yielded significant increase in accuracy level.[3].The paper[4] aims at the problem involved in automatically normalizing social media English and it provides the performance evaluation of two leading open source spell checkers on the data taken micro blogging service of twitter and measured its accuracy. Basheer Hawwash et al.

developed a Stream –Dashboard framework to support detecting and tracking evolving discussion clusters in Twitter.Their results show an ability to automatically detecttrending stories and their major milestones, such as the start, end, and major intermediateevents of the story [5].Ratab Gull et al.discusses on the method to smooth and ease the task of opinion mining with help of linguistic analysis and opinion classifier for determining sentiments on the political parties in Pakistan[6].Tajinder Singh et al proposed a method in which n-gram feature was used to find the bindings and conditionalrandom fields to check the significance of slang word which clearly indicated an improvement in classification accuracy[7].Sanket Patil et al states that Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entities in text into pre-defined categories.[8] shows that high accuracy in named entity recognition is achieved by applying the method Stanford NER.

III. METHODOLOGY

Social media is nothing but a bundle of websites where users network socially. To be characterized as social media a website should pose a the following charateristics.1.should provide web space for the user to upload content 2.The user are given a unique web identity.3.The users are asked to build their profile 4.users are requested to connect with friends 5.Users are allowed to post contents in real time. 6.Members are given rights to comment on the posts of others 7.All posts are time stamped.Understanding and pre- processing such type of social media data to glean actionable and interesting patterns presents lots of challenges and opportunities for researchers.

Pre-processing of text includes mechanism such as extraction, Data Transformation and data creation

A. Data extraction

The process of extraction involves extracting the short text from the text along with date and time of texts. To enable the extraction, necessary scripting language was used to extract the data.

B. Data Transformation

The raw text is transformed intostructured onethrough text cleaning. The various process involved in text cleaning are 1.Converting all short texts into Lower case 2.Removing emoticons and punctuation 3.Removing URL's. Towards Text parsing: The extracted texts are then parsed using the steps stated 1.Extracting Hash Tags 2.Extracting timestamps 3.Extract direction change indicators 4.Extract Locations

C. *Data Creation*

Once the transformation phase is over, the data is restructured into the database as Raw text, Hash tags, Noun entity, timestamp, Direction change indicators, Hash tag buckets, location etc as Figure .2.The Raw tweet is nothing but the exact text obtained from social media. The short text is the one that is cleaned from all punctuations, emoticons, numbers etc and the alphabets are converted into lowercase as shown in Figure .1. The Hash tag column stores all the highlights by the user and the once appearing with the symbol '#'.The noun entity column holdsall the nouns and the timestamp entity stores the date and time of the short text. Direction indicators are the one which change the whole context of the text such as 'and', 'or'. Hash tag buckets keep record of the quantity and impact of hash tags in the form of numbers. The location column stores the location of the user who texted .

Figure 1. Transformation of Raw Text

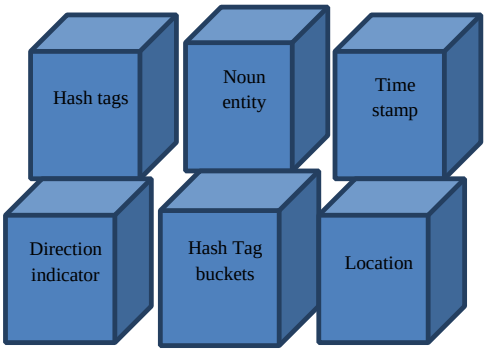

Figure 2. Structure of the data base for transformed information

The transformed information is subjected to online text cleaning, white space removal, expanding abbreviation, stop words removal,stemming, negation handling and finally feature selection[15]. Feature selection is a preprocessing techniques that reduces the data dimensionality, thereby chooses features that are important for prediction.The popular feature selection method are Term Frequency Inverse Document Frequency (TF-IDF),feature frequency (FF), and feature presence(FP).The number of occurrences in the document is denoted by (FF). TF-IDF formulais given by

$$TF\text{-}IDF = FF * Log((N/DF)$$

IV. EXPERIMENTAL ANALYSIS AND RESULTS

The dataset contains14000 tweets from Stanford Twitter Sentiment Corpus[10][11]. The tweets contains both positive and negative emotions .To preprocess the datasets we have followed the preprocessing techniques such as online text

cleaning, white space removal ,removed nonEnglish word,stop words ,abbreviations ,stemming and replaced the links and URL on tokens"http" and finally feature selection using TF-IDF[12][13] . The processed data is ready for classification process .*Figure .3 and Table .1*clearly states the comparison between the processed and Un-processed data taking into account 1996 attributes .

TABLE .1 COMPARISON BETWEEN UNPROCESSED AND PROCESSED DATA

Attributes 1966	Correct English	English with errors
Unprocessed data	1671.1	294.9
Preprocessed data	1867.7	98.3

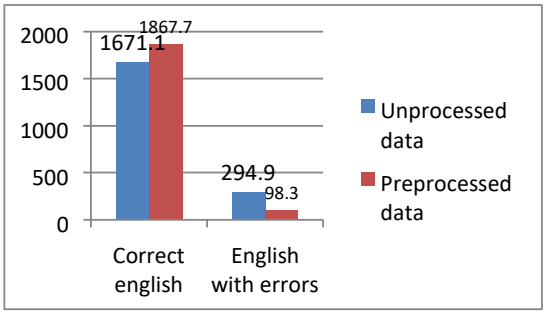

Figure .3 Analysis between unprocessed and processed data

In order to evaluate the measure of the processed data, sentiment classification task was performed using Naïve Bayes and SVM classifier. The results obtained using sentiment analysis are tabulated in*Table .2*

TABLE II. PERFORMANCE OF CLASSIFIERS

Classifier		Neg F	Pos F
Naïve Bayes	Unigram	0.53	0.50
	Bigram	0.56	0.40
SVM	Unigram	0.61	0.59
	Bigram	0.45	0.53

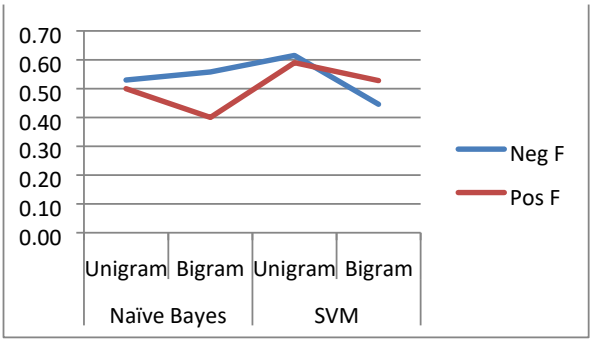

Figure .4 F-measure for Naïve Bayes and SVM

From *Figure .4* it clearly states that the F-score for SVM classifier is greater when compared to Naïve Bayes Classifier and SVM classifier outperforms in the context of social media mining.

V. Conclusion And Future Works

This work is to analyze the importance of preprocessing and normalizing the social messages obtained from social networks .The proposed work takes into comparison both the processed and unprocessed twitterdata from Stanford twitter sentiment corpus and also the influence of two popular classification technique such as Naïve Bayes and SVM on the preprocessed data[9]. The summary states that SVM classificationobtained higher F-score when compared to other. The future work could be devising a complete framework by incorporating and integrating open source parser,spell checker, dictionaries, etc into a framework system [14].

Reference

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

[2] GiulioAngiani, Laura Ferrari, TomasoFontanini, Paolo Fornacciari, EleonoraIotti, Federico Magliani, and Stefano Manicardi, A Comparison between PreprocessingTechniques for Sentiment Analysis in Twitter

[3] Emma Haddia, XiaohuiLiua, Yong Shib "The Role of Text Preprocessing in Sentiment Analysis",Procedia Computer Science 17 (2013) 26 – 32

[4] Eleanor Clarka* and Kenji Arakia, "Text normalization in social media: progress, problems and applications for a pre-processing system of casual English", Procedia - Social and Behavioral Sciences 27 (2011) 2 – 11

[5] BasheerHawwash, OlfaNasraoui," From Tweets to Stories: Using Stream-Dashboard to weave the twitter data stream into dynamic cluster models", JMLR: Workshop and Conference Proceedings 36:182–197, 2014

[6] RatabGulla*, Umar Shoaiba, Saba Rasheedb, WashmaAbidb,BeenishZahoorb, " Pre Processing of Twitter's Data for Opinion Mining in PoliticalContext, Procedia Computer Science 96 (2016) 1560 – 1570
[7] TajinderSingh ,MadhuKumari, "Role of Text Pre-Processing in Twitter Sentiment Analysis",Procedia Computer Science 89 (2016) 549 – 554
[8] SanketPatil, Prof. VarshaWangikar, Prof. K. Jayamalini, " Tweet data Preprocessing and Segmentation to NER" , International Journal of Scientific & Engineering Research, Volume 8, Issue 1, January-2017 2075 ISSN 2229-5518.
[9] Nimala.K.., Magesh S., ThamizhArasan. R., " Performance Analysis of Machine Learning Classifiers for Sentiment Analysis on Social Media Datasets" International journal of pure and Applied Mathematics,Volume 115,No.6,2017,597-603
[10] http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip [11] http://twittercommunity.com
[12] http://www.kdnuggets.com/2016/06/mining-twitter-data-python-part-2.html
[13] https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0
[14] http://cse.iitkgp.ac.in/~pawang/courses/SC16/Social_media_analytics.pdf
[15] https://dazeinfo.com/2015/04/14/dirty-social-media-data-ismisguiding-brands-tracking-consumer-behaviour-report/-Dirty