Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 165 (2019) 245–251

Procedia Computer Science

www.elsevier.com/locate/procedia

INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019

Analysis of Twitter Specific Preprocessing Technique for Tweets

Dharini Ramachandran*, Parvathi R

Vellore Institute of Technology, Chennai, 600127, India

Abstract

Social media plays an important role in capturing the thoughts of people in their own representation of sentences. Twitter, the popular microblog, traps abundant rich information in it, in terms of short texts posted by users. Processing of such colloquial and informal sentences require a specific preprocessing technique that can alleviate its understanding and analysis. Understanding these natural language sentences to produce a desired result is a challenging task. The paper aims at analyzing, identifying and explaining the advantages of twitter specific preprocessing technique and find out if it performs equally well with the established baseline preprocessing method. Experiment is carried out for classifying tweets to assess the performance of preprocessing techniques. The results indicate the effectiveness of twitter specific preprocessing technique and its competence with the baseline text preprocessing technique.

Keywords: Preprocessing; Twitter analysis; Tweet NLP; Social Media Text Analytics; NLTK preprocessing

1. Introduction

A natural language, to be understood by the computer and produce some desired result is a challenging task. One of the main reasons is the possibility of multiple interpretations for a same sentence. The metaphorical interpretations are hard to be learned by the computers. The ambiguities in words, sentences, part of speech, syntax, meaning and much more make the understanding of the text much difficult. Identifying sarcasm and humor from complex sentences, understanding the semantics of the text are all tough tasks in natural language processing. The digital text prevailing on social media gives abundant rich information from which many crucial intelligence can be uncovered. Some of the popular applications of text analytics on social media are News Detection, Event detection, Sentiment

*Corresponding author. Tel.: +91-979087259
E-mail address: dharini.r2014@vit.ac.in

10.1016/j.procs.2020.01.083

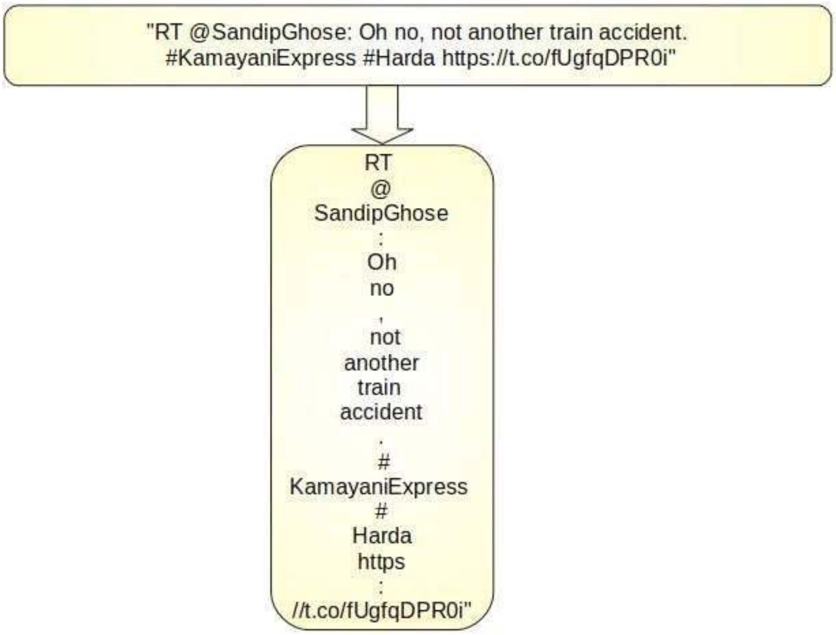

"RT @SandipGhose: Oh no, not another train accident.
#KamayaniExpress #Harda https://t.co/fUgfqDPR0i"
RT
@
SandipGhose
:
Oh
no
,
not
another
train
accident
.
#
KamayaniExpress
#
Harda
https
:
//t.co/fUgfqDPR0i"

Figure 1: Tokenization

analysis, Question answering system and Social Tagging. Microblogs are social media platforms that aid people to post messages in short texts and share it with the world. This kind of texts portray the mindset of the person who has written the message, which is helpful in many ways. The content from microblogs are analyzed to understand the personal views of the person. In application like situational awareness during disasters, the microblog texts aids us with the right direction to be aware of the situation and needs of the people under threat. One of the popular microblogs in Twitter, which allows its users to send short texts of 280 characters (tweets). The easy availability of twitter (on mobile phone, tablets, computers) in many devices and easy to use interface has attracted many users. Twitter is also welcomed by the research community for the availability of its API (Application Programming Interface) to extract the data. A tweet consists of messages posted by the user that may include the username with a '@' symbol in the front, the topic about which the tweet is about with a '#' (hashtag) symbol, the retweet with a 'RT' and much more. The first step in natural language processing is the preprocessing performed on the texts. The aim of the paper is to understand the advantages of Twitter specific preprocessing methods to preprocess the tweets rather than using normal preprocessing techniques for the same.

2. Background

The tweets posted by users contains texts written in informal manner with slangs, abbreviations, emoticons, URLs and Hashtags [1]. The noise, the tweets with respect to different geographic location, age, gender and much more are also prevailing challenges in processing the tweets [2]. The traditional preprocessing techniques that were trained on news texts, perform poorly for this kind of tweet texts [3]. Hence, a preprocessing technique that is trained for the informal nature of tweets has become a necessity. A work on such technique is started in [4], in which a Part=Of-Speech tagger is released for texts that are specific to Twitter. Later many improvements are performed and the new version is released [5]. The authors made the code available for research use in [6]. The package consists of a tokenizer, a tagger, a dependency parser [7] and a hierarchical word cluster.

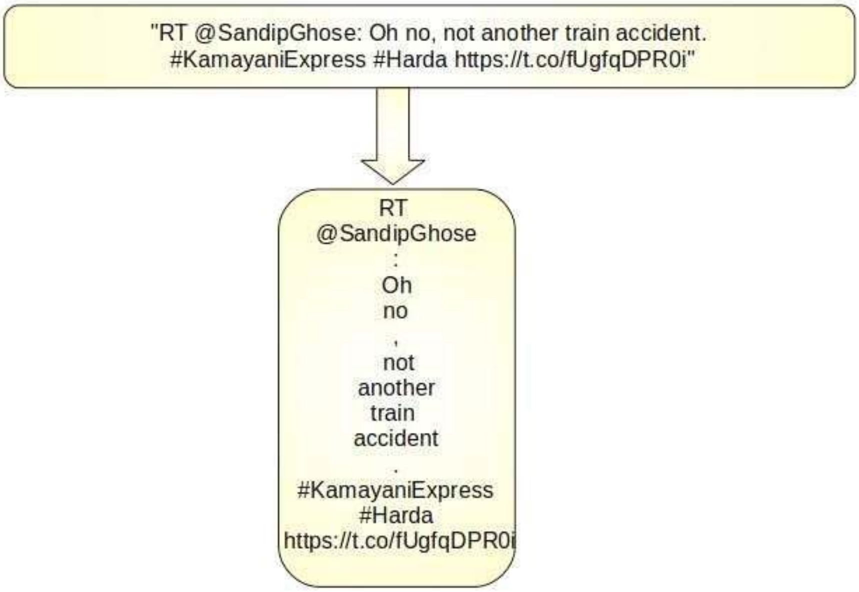

"RT @SandipGhose: Oh no, not another train accident. #KamayaniExpress #Harda https://t.co/fUgfqDPR0i"
RT
@SandipGhose
:
Oh
no
,
not
another
train
accident
.
#KamayaniExpress
#Harda
https://t.co/fUgfqDPR0i

Figure 2: Twitter Specific Tokenization

3. Preprocessing

Preprocessing is a crucial step in the processing of text. A text can comprise of words, sentences and paragraphs. A meaningful sequence of characters is considered as text. To feed the text data to a machine learning algorithm, in a better form rather than in its natural form, the preprocessing techniques are used.

3.1. Text Preprocessing

The most widely used text preprocessing techniques are available in the NLTK (Natural Language Tool Kit) [8] library in python. The tokenization is performed based on the regular expressions to tokenize them as in Penn Treebank. The Penn Treebank tagset is used for the Part-Of-Speech tagging in NLTK.

3.2. Twitter Specific Preprocessing

The technique introduced in [6] essentially deals with the tweets and preprocesses them. From the team of Carnegie Mellon University, a library named TweetNLP is released which includes a Tokenizer and a Part-Of-Speech tagger, along with many other resources. A new tagset is released for the English language tweets. The annotation guidelines are explained in the paper [5].

4. Experimental Details

The aim of the experiment is to compare and analyze the impact of the two different preprocessing techniques used. The first step is to extract tweets from Twitter that are posted during disasters and the next step is to preprocess them. In preprocessing, we use two different preprocessing techniques and intend to analyze them. The techniques consist of stages such as tokenization, Part-Of-Speech tagging and stopword removal. At the end, the preprocessed tweets are fed to machine learning algorithms for classification into two categories namely Disaster and Non Disaster. The performance is evaluated based on the accuracy of the outcome. The results are analyzed to understand the working of the preprocessing techniques.

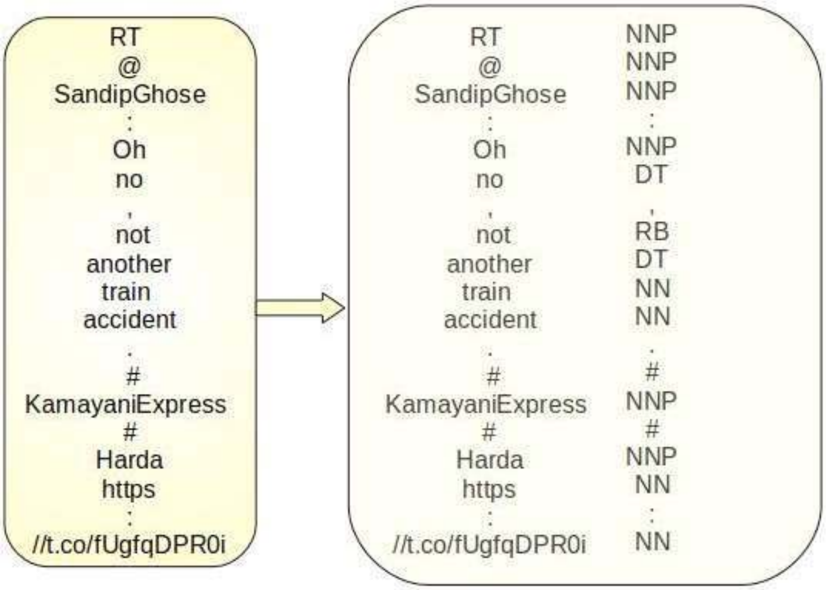


Figure 3: Part-Of-Speech Tagging

4.1. *Tweet Extraction*

The tweets are extracted from twitter using Twitter API using Tweepy [9] library. Every tweet in twitter consists of an identifier and these tweet ids are utilized to extract them. The tweet ids are obtained from the work published in [10] and the extracted tweets are stored in MongoDB. The tweets are extracted for two classes - Disaster and Non Disaster.

4.2. *Preprocessing*

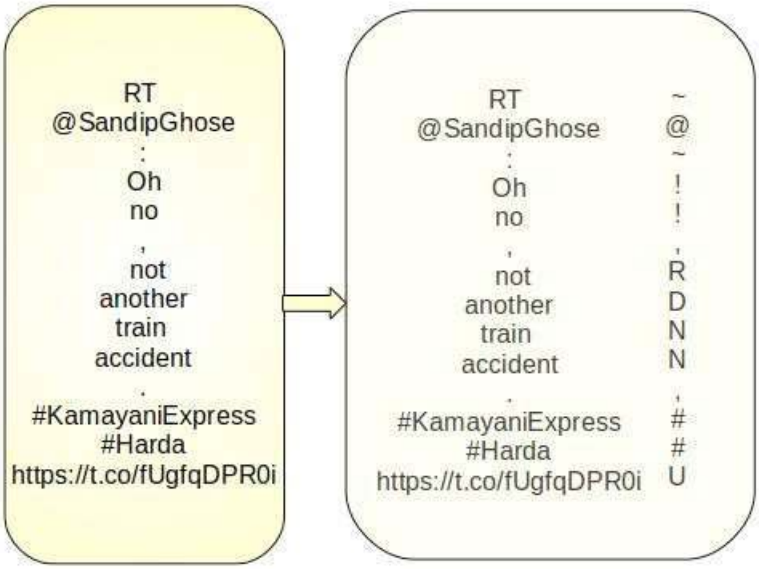
The tweets are preprocessed using normal text preprocessing technique and twitter specific preprocessing technique. Both the techniques go through the following stages to complete the preprocessing. The normal preprocessing technique is implemented using NLTK (Natural Language Tool Kit) library [8], the twitter specific preprocessing is implemented using TweetArc NLP [6].

4.2.1. *Tokenization*

To process the text, it is important to find the boundaries of words. The boundaries are identified using the spaces and punctuations. This process of splitting a sentence into meaningful parts and identifying the individual entities in the sentence is named as Tokenization. The figure1shows the tokenization performed with NLTK library. The figure 2shows the tokenization performed with TweetArc NLP. The main advantage of the twitter specific tokenization is the complete separation of URLs and Hashtags present in the tweet. When observed in the normal technique, the URLs are tokenized into many parts and the '#' is separated from its word. The identification of hashtags is immensely helpful in further processing in many applications such as trend detection, opinion mining, event detection and much more.

4.2.2. *Part-Of-Speech Tagging*

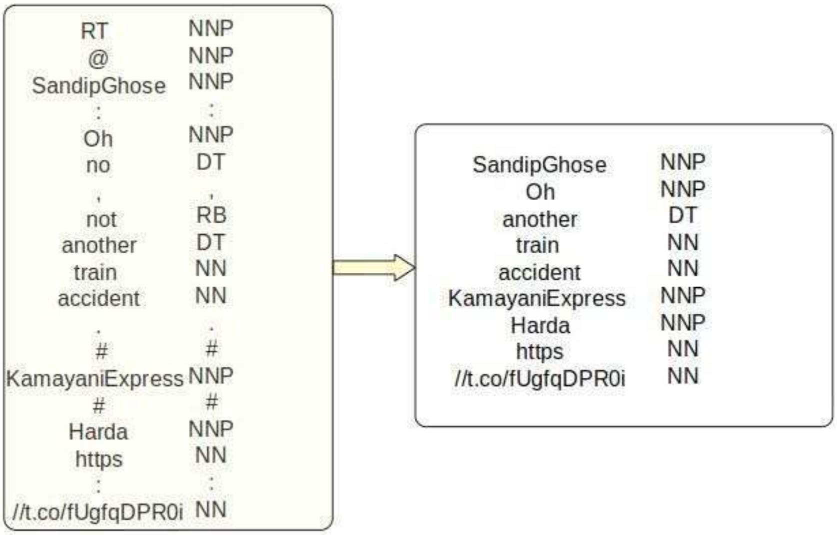
The main part of the preprocessing is the Part-Of-Speech (POS) tagging.The POS tagging identifies the word class based on the position of the word in the sentence. As shown in the figures3and4the twitter specific technique has an advantage of identifying the URLs. Also the twitter specific preprocessing technique can identify the Emoticons,



RT
@SandipGhose
:
Oh
no
,
not
another
train
accident
.
#KamayaniExpress
#Harda
https://t.co/fUgfqDPR0i

RT	~
@SandipGhose	@
:	~
Oh	!
no	!
,	,
not	R
another	D
train	N
accident	N
.	,
#KamayaniExpress	#
#Harda	#
https://t.co/fUgfqDPR0i	U

Figure 4: Twitter specific Part-Of-Speech Tagging



RT	NNP
@	NNP
SandipGhose	NNP
:	:
Oh	NNP
no	DT
,	,
not	RB
another	DT
train	NN
accident	NN
.	.
#	#
KamayaniExpress	NNP
#	#
Harda	NNP
https	NN
:	:
//t.co/fUgfqDPR0i	NN

SandipGhose	NNP
Oh	NNP
another	DT
train	NN
accident	NN
KamayaniExpress	NNP
Harda	NNP
https	NN
//t.co/fUgfqDPR0i	NN

Figure 5: Stopword Removal

’@’ mentions (usernames), ’RT’ tags (retweets) which are used liberally in the tweets. The normal preprocessing technique does not consider the above mentioned tags and so they are tagged differently.

4.2.3. Stopword Removal

The stopwords are those words that are less helpful in the further analysis of the tweet. These words removed from the tweets before fed to the machine learning algorithm. The figure5shows how e ffective the stopword removal is due to the explicit tagging of twitter specific words.The number of words being sent for the analysis is much less compared to the normal preprocessing technique as it can be seen in figure6.

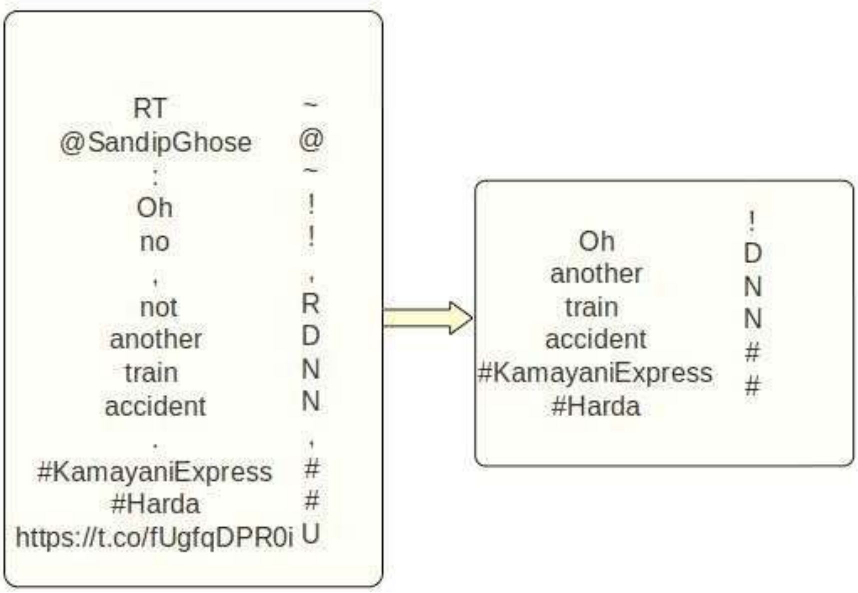


Figure 6: Twitter Specific Stopword Removal

Preprocessing \Classification		**Naive Bayes**	**Naive Bayes (Vector Representation)**
Text Preprocessing	Word and POS	0.89	0.79
	Word	0.92	0.81
Twitter Specific Preprocessing	Word and POS	0.91	0.79
	Word	0.91	0.79

Table 1: Comparison of Accuracy

4.3. Classification

The tweets are preprocessed and gathered in to a dataset with corresponding labels namely Disaster and Non Disaster. The Naive Bayes machine learning algorithm is utilized for classifying the tweets. In Naive bayes algorithm, the words in the tweets are fed as such to be classified and in another variation, the words in the tweets are converted into One-Hot encodings and then fed to the algorithm. The features utilized in the classification are the actual words, the words and POS tags. The results of the classification can be seen in table1

4.4. Result and Discussion

The results indicate that the twitter specific preprocessing technique performs competently with traditional text preprocessing technique as shown in figure7. The figure shows the accuracy scores of two preprocessing techniques, for both Naive Bayes (using actual word) and Naive Bayes with vector representation for word. It can be seen that the accuracy has improved in the 'Word and POS' feature category when using the Twitter Specific Preprocessing technique. In other categories also, the scores indicate that the two techniques are equally efficient.

The preprocessing technique implemented with the NLTK library is the one that is most widely used for preprocessing the English texts. To have another technique, with so many merits, that works equally well with this established one is a major welcome. The twitter specific techniques identify the emoticons in the tweet which is a major advantage in applications such as sentiment analysis and opinion mining. As tweets are short texts, with a fixed number of characters, people usually expresses their opinion in form of emoticons. Also, the ability to identify the hashtags is aids tremendously in identifying the topic of the tweet. This topic of the tweet is most required information in

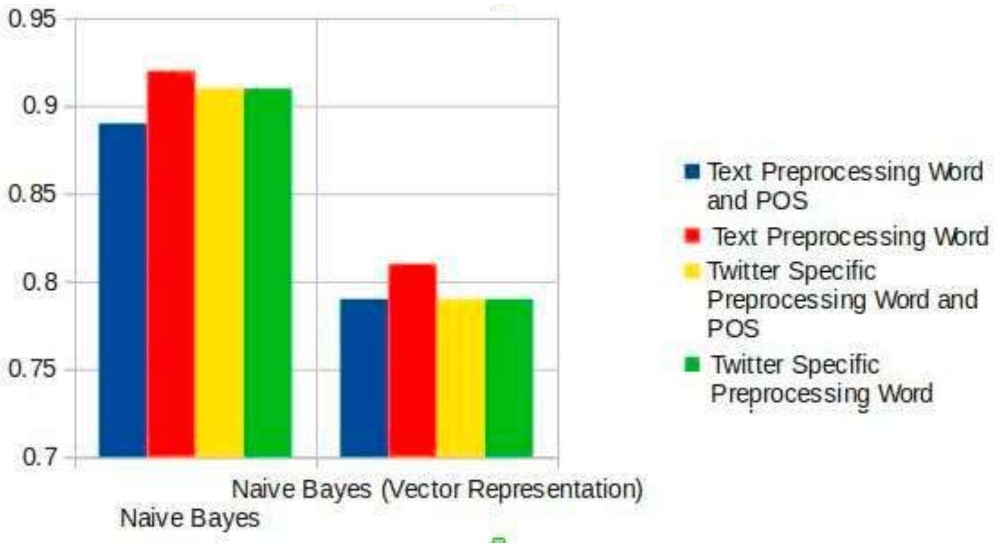


Figure 7: Comparison of Accuracy

tweet analysis and is useful in all application. The separation of URL is helpful in identifying the external source of information that backs up the tweets' credibility

5. Conclusion

The paper aims in finding out if the twitter specific preprocessing technique works efficiently as the broadly accepted traditional text preprocessing technique. The results of the experiment in terms of accuracy, indicate that both work equally competent. In all stages of preprocessing such as tokenization, Part-Of-Speech tagging and stopword removal, the twitter specific techniques show visible advantages. With comparable performance and multifold advantages, the twitter specific preprocessing technique can be used in applications which analyze the tweet contents explicitly.

References

[1] S. Carter, W. Weerkamp, M. Tsagkias, Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text, Language Resources and Evaluation 47 (1) (2013) 195–215.

[2] L. Derczynski, D. Maynard, G. Rizzo, M. Van Erp, G. Gorrell, R. Troncy, J. Petrak, K. Bontcheva, Analysis of named entity recognition and linking for tweets, Information Processing & Management 51 (2) (2015) 32–49.

[3] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, M. Dredze, Annotating named entities in twitter data with crowdsourcing, in: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Association for Computational Linguistics, 2010, pp. 80–88.

[4] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, N. A. Smith, Part-of-speech tagging for twitter: Annotation, features, and experiments, Tech. rep., Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science (2010).

[5] O. Owoputi, B. OConnor, C. Dyer, K. Gimpel, N. Schneider, Part-of-speech tagging for twitter: Word clusters and other advances.

[6] Tweet NLP, http://www.cs.cmu.edu/~ark/TweetNLP/.

[7] L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, N. A. Smith, A dependency parser for tweets, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1001–1012.

[8] NLTK library, https://www.nltk.org/index.html.

[9] Tweepy library, http://docs.tweepy.org/en/v3.5.0/.

[10] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, S. Ghosh, Extracting situational information from microblogs during disaster events: a classification-summarization approach, in: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, ACM, 2015, pp. 583–592.