TEST Engineering & Management

September-October 2019
Page No.17 - 26

Proposed Preprocessing Methods for Manipulate Text of Tweet

Hayder Mahmood Salman [1,a]
[1]Computer Science Department, Al-Turath University College, Baghdad, Iraq

Article Info
Page Number: 17 - 26
Publication Issue:
September-October 2019

Article History
Article Received: *08 January 2019*
Revised: *06 April 2019*
Accepted: *12 August 2019*
Publication: *20 October 2019*

Abstract: Social media provides plentiful information to study people's ideas and opinions about events in this world, such as political, economic issues, disasters and others. The tweets that published must study and classify. In a classification system for text mining, an important step is the preprocessing phase. It is often underestimated. This paper uses Twitter data as a case study. The aim of the paper is to propose methods for preprocessing for each text of tweet. In this paper the propose in preprocessing steps that include proposed manipulate Hashtag state method that used to extract another important word to assist in other processing processes and proposed enhancement stemming algorithm to return the entered word to its source without Affix, which are both Suffix and Prefix extensions and final results of the preprocessing process are presented based on number of words after preprocessing.

Keywords: *Social media, Text classification, Preprocessing, Twitter.*

1. Introduction:

Social media is a convenient space for people to clarification their opinions on certain events and communicate with each other. Twitter is a social networking application which allows people to micro-blog about a broad range of topics. A Tweet: is a text message containing no more than 140 characters to put down thoughts were people write what is happing or what they like to share. Thus, interested people can view and interact with it. In our research tweet present the main source for the classification [1].

Text classification is an embodiment and essence issue for various applications is, similar to sentiment analysis, smart replies or spam identification. It is an issue contemplated generally in recent years and different techniques used to tackle this issue. Text classification intends to dispense records to numerous or one classification. On the off chance that record distributed to in excess of one class, if report assigned to in excess of one class, it called "multi-label" and if an archive apportioned to just a single class, it called "single label" [2]. Most strategies rely to upon representing text as a text vector to order it. This text vector contains recurrence of each word in content. It can be more complex and speak to a few highlights that removed from the content [3].

The pre-processing is an important step to provide satisfactory results. The length of the Tweet text where only contains 140 characters or less. Tweets contain non-useful data in the process of categorizing text such as a URL (global web page address), a label, numbers, and stop words. For example, 'Hurricane Sandy! #Hurricane (Bonnier) http: // twittter.com'. It is important to remove these additions or manipulate them in tweets so as not to affect the classification process [4].

This research is designed to manipulate the texts of tweets on Twitter based on proposed manipulate hashtag state, remove additions, tokenization, remove of stop words,proposed stemming algorithm and lemmatization, where important words are extracted and the additives that affect the extraction process are removed

2. Related Works

In 2015 Liza Wikarsaet al [5]. Three main phases of the text mining utilized were preprocessing, processing, and validation. Activities conducted in the preprocessing phase were case folding, cleansing, stop-word removal, negation conversion, and tokenization to the training data and the test databased on the sentiment analysis that performed morphological analysis to build several models.In 2015 Zhao Jianqiang [6]. Discuss the effects of pre-processing on sentiment classification performance. Evaluated the effects of URL, stop word, repeated letters, negation, acronym and number on sentiment classification performance using two feature models and four classifiers on five Twitter datasets. In 2016Lizaet al [7].Several processes were done in the pre-processing phase, including removal of URLs, punctuation, and stop words, tokenization, and stemming. Later, the application automatically converted the pre-processed tweets into set of features vector using Bag of Words. In 2016 TajinderSingh et al [8]. The proposed method of pre-processing relies on the bindings of slang words on other coexisting words to check the significance and sentiment translation of the slang word. We have used n-gram to find the bindings and conditional random fields to check the significance of slang word. In 2017 Itisha et al [9]. explore process for preprocessing, the process which comprises of 2 segments: denoising such as removal of StopWords, URLs, username, punctuation etc. and normalization such as conversion of Non-standard words to their canonical forms. Tweets are pre-processed by normalizing elongated words, misspelled words, informal acronyms, negation handling, In 2017 Hussein K. Al-Khafaji et al. presents the design and implementation of a system for English tweets segmentation, cleaning, stop words removing, and stemming.

3. Data Collection phase

The first step for this work is a data collection. Basic data collect from twitter using a Twitter Application Programming Interface (API). The challenges in data collection are online connectivity and different data types when collecting data, including images, videos, location, date and text data. The important thing is to extract only tweet text. Data is collect based on the time period. Each record contains the date, location, timestamp, publisher, and text data, where text data is extract only and is grouped into a one dataset of Initial data. Set of 1000 tweets take from initial data chased as training data. This training data used to extract More than 50 words used frequently. Set of 1002 tweets take from initial data as testing data, where this test data divide into five datasets to process and classify it. Figure (1) show the division of the data that is collected.

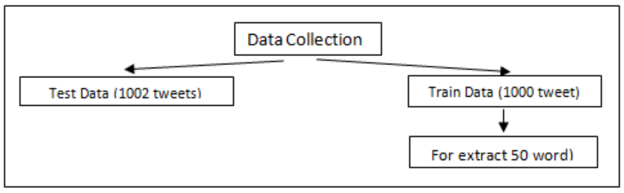


Figure (1) division of the data collected

4. Data preprocessing phase

The pre-processing step is an important step in the proposed system to provide satisfactory results. The length of the Tweet text where only contains 140 characters or less. Tweets contain non-useful data in the process of categorizing text such as a URL (global web page address), a label, numbers, and stop words. For example, 'Hurricane Sandy! #Hurricane (Bonnier) http: // twittter.com'. It is important to remove these additions or manipulate them in tweets so as not to affect the classification process. There are many internal processes in this step such as manipulate hashtag, remove URL, remove special character, remove additions, tokenization, remove stop words, stemming, lemmatization and Part of Speech (POS). Collected and divided tweets is the inputs to the preprocessing step and the output of this step is a series of important words that used in the feature extraction.

4.1 Proposed manipulate Hashtag state

The hashtag is a keyword begin with the hash mark (#) and set of characters, for instance #HurricaneSandy. It is a feature in twitter used to facilitate effective search and it is possible to extract important words from this Hashtag. Manipulate of hashtag state step based on utilize pattern matching of hashtag keyword with the list of more 50 words used frequently in relevant training data for chosen event. Algorithm (1) show manipulate hashtag state.

Algorithm (1) Manipulate Hashtag state

```
Input: Tweets, List (X) of more 50 words used frequently in event, empty string (H)
Output : string (tweet) after manipulate hashtags
Step1:For each word in Tweet  do
        -   If the word starts with  "#"
                Put word in string (tweet)
Step 2: Delete "#"
        -   Read first character in keyword index (i=0)
Step 3: While not read space " " do
        -   Delete any character in string (H)
        -   Put it in string (H)
        -   If string (H) match any word in list (X)
                Put in string (tweet)
        -   Else
            Readindex(i=i+1) from keyword of hashtag
            Combine index (i) + index (i+1)
        -   End while
Step 4: Return string (tweet) after manipulate hashtags

Step5: End For
```

4.2 Remove additions

During the classification process, not important data in tweets affect to the classification process. For instance, URL (sUniform Resource Locator), a label, numbers and special character. URL used to specify the addresses on the World Wide Web. For example, "Hurricane Sandy! 2012 (Bonnier) HTTP: // twittter.com". It is better to remove these additions or manipulate them in tweets not affect on the classification process. This work used pattern matching to eliminate these additives. For example, a URL with a static pattern starting with "HTTP: //" or " HTTPS: //" will be deleted when it is found and delete numbers, label and Special characters as in the same way.

4.3 Tokenization

In this phase, each text document is partitioned into a number of tokens. The text is converting into a list of words. For example the text "Hurricane Sandy devastated cities" when pass in this phase the output is : ['Hurricane', 'Sandy', 'devastated', 'cities']. Algorithm (2) Show the steps of tokenization process.

```
Algorithm (2) Tokenization process
Input: Text of tweet
Output : List  of tokens
Step1: while not end of file do
        -  For each Tweet  do
             •  Split the Tweet to token when read: white space, solidus, stroke, semicolon,
                comma or dot.
             •  Put token in tokens List
        Endfor
Step 2: Return List of tokens
```

4.4Stop Word Recognition and Eliminator

The most important step in preprocessing is recognize and eliminate of the stop words by matching every word in text with stop words list, these words are the most frequently used in English language. Table (1) shows the stop words list. Algorithm (3) show the steps of remove the stop words.

Table (1): Sample of stop words

a's	able	about	above	appropriate	are	aren't	around
according	couldn't	first	couldn't	couldn't	couldn't	ask	asking
almost	did	for	few	he	with	with	been
also	although	always	am	has	him	behind	being
among	have	an	and	believe	below	beside	besides
anyone	it	want	anyway	both	brief	but	by

```
Algorithm (3) remove stop words
Input: The output of algorithm 2
          Stop words list
Output :List oftokens without stopword
Step1: While not end of file Do
          -  For each token in tweet
          -  Match token with stopwords list
          -  If token in stop word list
                Remove it
          -  Else
                 Put token in list of token
          End for
          -  End while
Step2: ReturnList oftokens without stop word
```

4.5 Proposed Enhancement Stemming algorithm

The stemming process work to return the entered word to its source without Affix which are both Suffix and Prefix extensions, in addition to setting the grammar rules on each word. Where a number of rules and conditions are applied to ensure retrieving the original word. The proposed method is applied rules on each word when number of letters is more than two characters. As an example, the derived words introduction, introducing, introduces are converted into stem word introduce. Algorithm (4) show stemming algorithm.

Algorithm (4) Stemming algorithm

Input : output of algorithm (3)
Output : stem of all input tokens

```
Step1:For each token in list of token
        -   If token contain of two letters or minimal then:
        •   don't change it
        Else
        •   Go to step 2
Step 2:
        -   If the token ends with "sses"
                •   Convert "sses" to "ss"
        -   If the token ends with "ss" or "us"
                •   Don't do any thing
        -   If token end with "ied" or "ies" and suffixes preceded by more than one letter
                •   Convert "ied" or "ies" to "i"
        -   If token end with "ied" or "ies" and suffixes not preceded by more than one letter
                •   Convert "ied" or "ies" to "ie"
Step 3:
        -   if token ends with "at", "iz" or " bl"
                •   add a letter "e" after this suffixed ( luxuriat convert to luxuriate)
        -   if token ends with a double letter
                •   remove last letter in token (hopp convert to hop)
        -   if token ends with a double letter and the length of word is short
                •   add the letter "e" after double letter ( hop convert to hope)
        -   if the token end with "eed"," eedly", " edly", "ed ", "ingly" ," ing", andpreceding the word
            part contains a vowel letter
                •   Convert "eed"," eedly" to " ee"
                •   delete "ed "," edly"," ing"," ingly"
Step 4:
        -   if suffixes preceded by a non-vowel which is not first letter of the word
                •   replace yor Y to " i" ( "by" Convert to "by","say" to "say")
Step 5:
    -   If the token ends with "tional", "enci", "anci", "abli", " ization", "izer", "ational", " ator", "
        aliti", " ation", "alism", " alli", "fulness", "ousli","ousness", "iveness", " bli" " iviti" ,
        "iveness"," iviti", "biliti",
                • "tional"   convert to " tion
                • "enci"   convert to  "ence"
                • "anci"  convert to" ance"
                • "abli"  convert to" able"
                • "izer"," ization"  convert to" ize"
                • " ator", " ation", "ational", , convert to "ate"
                • "alism"," alli" ," aliti", convert to "al"
                • "fulness"   convert to "ful"
                • "ousness", ousli" convert to "ous"
                • " iviti" , "iveness" convert to" ive"
                • " bli", "biliti"   convert to "ble"
    -   if preceded by "i"
                • "ogi"  convert to "og"
    -   If token ends with "li":   delete
                • "fulli"   Convert to" ful"
                • "lessli"   Convert to "less"
Step 6:
    -   If token ends with this suffixes do "tional", "ational", "alize", "l", "ful"," ness"
                • "tional"   Convert to "tion"
                • "ational"  Convert to "ate"
                • "alize" Convert to "al"
                • "icate", " ical", " iciti",  Convert to" ic"
                • " ness" , "ful", delete
Step 7 :Delet this suffixes:
                • "al"," ance", " er", " able"," ant", " ence", " ible" " ic" , " ize", "ment", " iti", "ate", "
                  ive", " ous", "  ent", Delete
    -   "ion"    delete if preceded by" s" or" t"
Step 8:
        find this suffixes, and do the conduct indicated
    -   "e"    delete
    -   "l"   delete and preceded by " l"
Step 9: End
```

4.6 Lemmatization

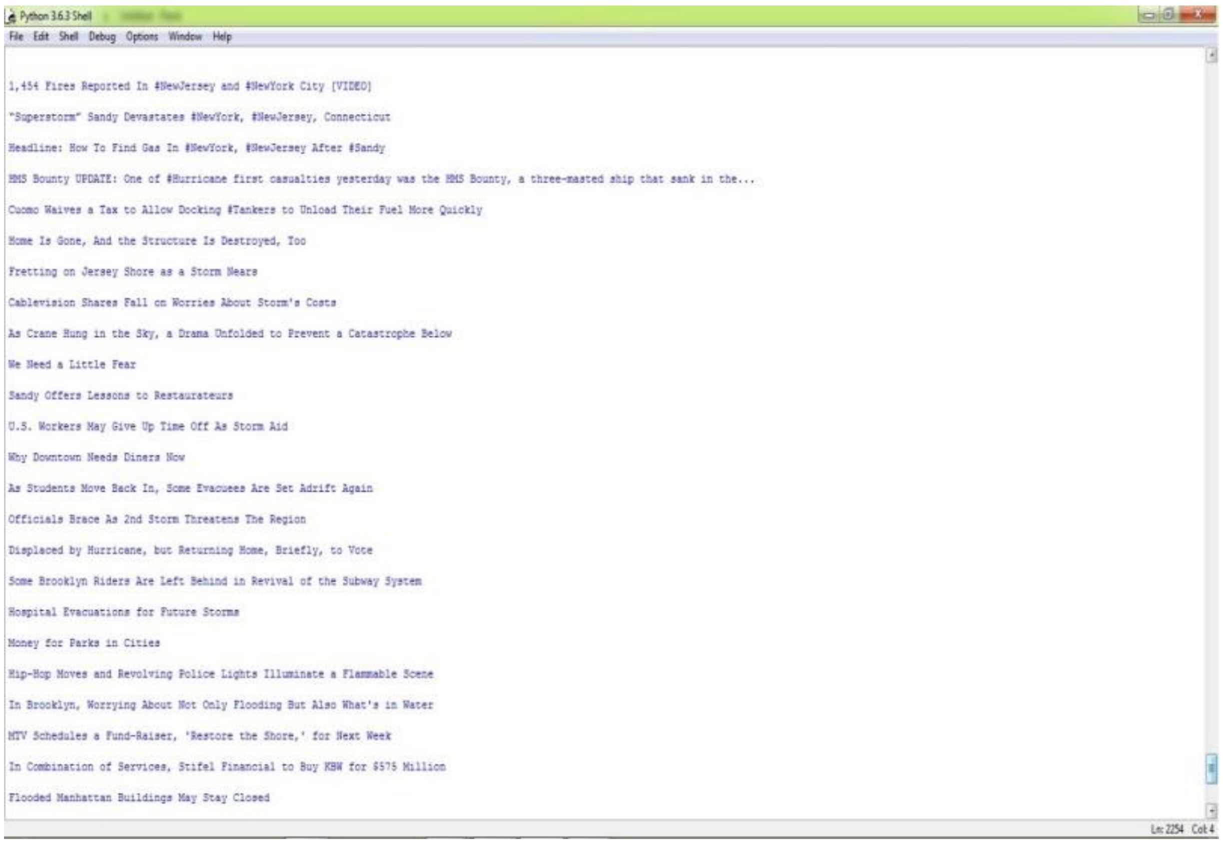
Lemmatization is a process of converting a words of a sentence to its dictionary form, where it returning different forms of the single word to its root. For examble, suffixes of words working, works and worked will change to get normalized form standing for the infinitive is work. Sometimes the normalized form may be different from the stem of the word. For example, the words compute, computing, computed would stem to compute, but their normalized form is infinitive of the verb compute. In general, many algorithms work on lemmatization words. The method used WordNet Lemmatizer algorithm that applied rules on each word when a number of letters is more than two characters. WordNet is a large lexical database that Contains verbs, adverbs, nouns and adjectives, where this part of speech are grouped in sets of cognitive synonyms (synsets). Synsets are connected by means of conceptual lexical and semantic relations. Table (2) show simple example of preprocessing phases.

Table (2) Example of preprocessing phases

Preprocessingphase	
Tweet	1454 Fires Reported In #NewJersey and #NewYork City
manipulate Hashtag state	1454 Fires Reported In New Jersey and New York City
Remove additions	Fires Reported In New Jersey and New York City
Tokenization	[Fires, Reported, In, New, Jersey, and, New, York, City]
Stop Words Removing	[Fires, Reported, New, Jersey, New, York, City]
Stemming	[Fire, Report, New, Jersey, New, York, Citi]
Lemmatization	[fire, report, new, jersey, new, york, City]

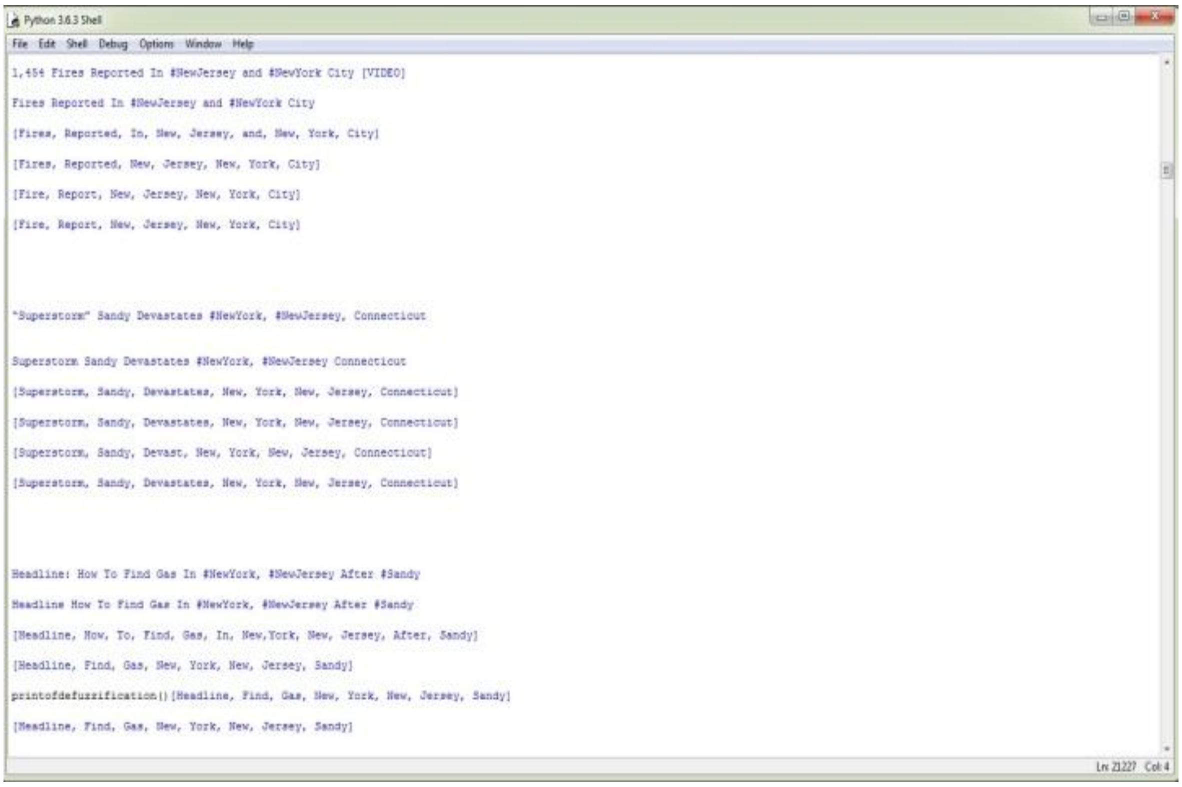
5 EXPERIMENTAL RESULTS

This section discusses the results of this paper. The data set is composed of the training and test sets. Training data contains 1000 tweeting and the test data contains 1002 tweets divided into five data sets. Table 2 show training and test dataset. Figure (2) show the snapshot of tweets before preprocessing. Figure (3) and figure (4) show the snapshot for the tweets after preprocessing.



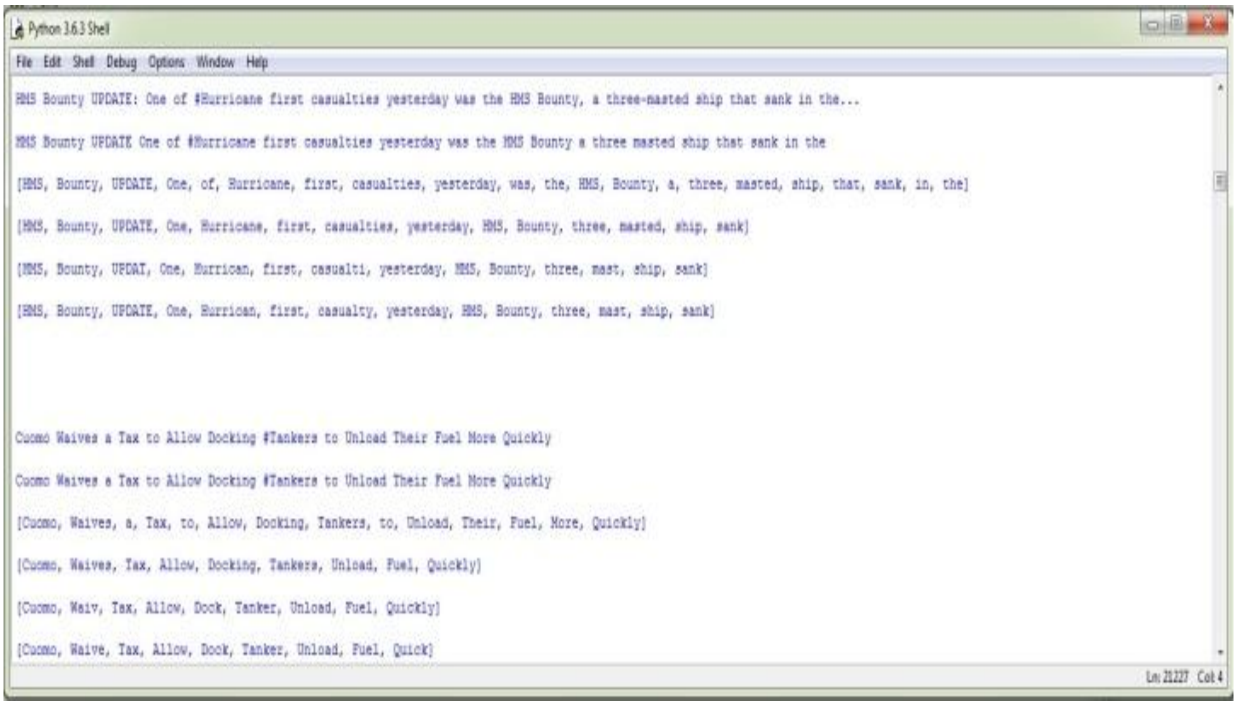
```
1,454 Fires Reported In #NewJersey and #NewYork City [VIDEO]

"Superstorm" Sandy Devastates #NewYork, #NewJersey, Connecticut

Headline: How To Find Gas In #NewYork, #NewJersey After #Sandy

HMS Bounty UPDATE: One of #Hurricane first casualties yesterday was the HMS Bounty, a three-masted ship that sank in the...

Cuomo Waives a Tax to Allow Docking #Tankers to Unload Their Fuel More Quickly

Home Is Gone, And the Structure Is Destroyed, Too

Fretting on Jersey Shore as a Storm Nears

Cablevision Shares Fall on Worries About Storm's Costs

As Crane Hung in the Sky, a Drama Unfolded to Prevent a Catastrophe Below

We Need a Little Fear

Sandy Offers Lessons to Restaurateurs

U.S. Workers May Give Up Time Off As Storm Aid

Why Downtown Needs Diners Now

As Students Move Back In, Some Evacuees Are Set Adrift Again

Officials Brace As 2nd Storm Threatens The Region

Displaced by Hurricane, but Returning Home, Briefly, to Vote

Some Brooklyn Riders Are Left Behind in Revival of the Subway System

Hospital Evacuations for Future Storms

Money for Parks in Cities

Hip-Hop Moves and Revolving Police Lights Illuminate a Flammable Scene

In Brooklyn, Worrying About Not Only Flooding But Also What's in Water

MTV Schedules a Fund-Raiser, 'Restore the Shore,' for Next Week

In Combination of Services, Stifel Financial to Buy KBW for $575 Million

Flooded Manhattan Buildings May Stay Closed
```

Figure (2) snapshot of tweets before preprocessing



```
1,454 Fires Reported In #NewJersey and #NewYork City [VIDEO]
Fires Reported In #NewJersey and #NewYork City
[Fires, Reported, In, New, Jersey, and, New, York, City]
[Fires, Reported, New, Jersey, New, York, City]
[Fire, Report, New, Jersey, New, York, City]
[Fire, Report, New, Jersey, New, York, City]



"Superstorm" Sandy Devastates #NewYork, #NewJersey, Connecticut

Superstorm Sandy Devastates #NewYork, #NewJersey Connecticut
[Superstorm, Sandy, Devastates, New, York, New, Jersey, Connecticut]
[Superstorm, Sandy, Devastates, New, York, New, Jersey, Connecticut]
[Superstorm, Sandy, Devast, New, York, New, Jersey, Connecticut]
[Superstorm, Sandy, Devastates, New, York, New, Jersey, Connecticut]



Headline: How To Find Gas In #NewYork, #NewJersey After #Sandy
Headline How To Find Gas In #NewYork, #NewJersey After #Sandy
[Headline, How, To, Find, Gas, In, New,York, New, Jersey, After, Sandy]
[Headline, Find, Gas, New, York, New, Jersey, Sandy]
printofdefuzzification()[Headline, Find, Gas, New, York, New, Jersey, Sandy]
[Headline, Find, Gas, New, York, New, Jersey, Sandy]
```

Figure (3) snapshot of the tweets after preprocessing



```
HMS Bounty UPDATE: One of #Hurricane first casualties yesterday was the HMS Bounty, a three-masted ship that sank in the...
HMS Bounty UPDATE One of #Hurricane first casualties yesterday was the HMS Bounty a three masted ship that sank in the
[HMS, Bounty, UPDATE, One, of, Hurricane, first, casualties, yesterday, was, the, HMS, Bounty, a, three, masted, ship, that, sank, in, the]
[HMS, Bounty, UPDATE, One, Hurricane, first, casualties, yesterday, HMS, Bounty, three, masted, ship, sank]
[HMS, Bounty, UPDAT, One, Hurrican, first, casualti, yesterday, HMS, Bounty, three, mast, ship, sank]
[HMS, Bounty, UPDATE, One, Hurrican, first, casualty, yesterday, HMS, Bounty, three, mast, ship, sank]



Cuomo Waives a Tax to Allow Docking #Tankers to Unload Their Fuel More Quickly
Cuomo Waives a Tax to Allow Docking #Tankers to Unload Their Fuel More Quickly
[Cuomo, Waives, a, Tax, to, Allow, Docking, Tankers, to, Unload, Their, Fuel, More, Quickly]
[Cuomo, Waives, Tax, Allow, Docking, Tankers, Unload, Fuel, Quickly]
[Cuomo, Waiv, Tax, Allow, Dock, Tanker, Unload, Fuel, Quickly]
[Cuomo, Waive, Tax, Allow, Dock, Tanker, Unload, Fuel, Quick]
```

Figure (4) snapshot of the tweets after preprocessing

The final results were calculated based on the number of words after each stage of pretreatment. Table (3) show Training and test dataset. Table (4) show the number of words before and after preprocessing. Table (5) show Detailed results for each phase of preprocessing.

Table (3) Training and test dataset

Dataset	Number of tweets
Training	1000
Dataset 1	216
Dataset 2	227
Dataset 3	175
Dataset 4	170
Dataset 5	214

Table (4) number of words before and after preprocessing

Dataset	Number of tweets	Number of words before preprocessing	Number of words after preprocessing
Training	1000	-	-
Dataset 1	216	1935	803
Dataset 2	227	2208	1129
Dataset 3	175	1742	912
Dataset 4	170	1866	892
Dataset 5	214	1837	837

Table (5) Detailed results for each phase of preprocessing

Dataset	Number of words before preprocessing	Number of words after manipulate Hashtag state	Number of words after Remove additions	Number of words after Tokenization	Number of words after remove Stop Word	Number of words after stemming	Number of words after lemmatization
Data set 1	1935	2095	1447	1447	803	803	803
Dataset 2	2208	2378	1697	1697	1129	1129	1129
Dataset 3	1742	1882	1357	1357	912	912	912
Dataset 4	1866	1991	1481	1481	892	892	892
Dataset 5	1837	2007	1365	1365	837	837	837

Through the table data details, the process of manipulate Hashtag state contributes to increase the number of important words that contribute to increase the accuracy of extract best feature from tweet. The process of Remove additions and remove stop words reduce the number of words. which the words and additions that affect the classification process are deleted. Table (6) shows the difference between this proposed methods and proposed preprocessing system [10].

The data set size (tweets number)	Words number after preprocessing based on Proposed Method	Words number after preprocessing based on preprocessing system
250000	71203	690640

details By observing the results of the table, the proposed method extracts more words than the preprocessing system [10]as a result of the reliance on the proposed methods, especially the manipulate of the hashtag state and remove additions states.

6 CONCLUSION

Preprocessing of text is a phase in all applications of the data mining. In text classification, it is used in all available research, where few works have been specifically set up to realize the position of each the basic preprocessing techniques that applied to textual data. This work based on a simple methodology. it applies each one of the process, Sequentially, to the raw data. it is worth noting that the use of proposed manipulate Hashtag state and proposed Enhancement Stemming algorithm enhance the performances in our tests. All other techniques provided improvements to the performances of classifier . This research based on data which originated from a Twitter. An analytical work should be performed on the different sets of data to have a comprehensive understanding of the different preprocessing process. The mix of of these process together is often correct. it should be better induced by empirical data and evaluations of result for different application domains and the peculiar nature of their textual data.

References:

[1] Nahon Karine and Crowston Kevin,"Introduction to the Digital and Social Media Track", IEEE, 49th Hawaii International Conference on System Sciences (HICSS), USA, 2016.

[2] Charu C. Aggarwal, "Data Classification Algorithms andApplications", Taylor & Francis Group, 2015.

[3] Rajni Jindal, Ruchika Malhotra, Abha Jain, "Techniques fortext classification: Literature review and current trends",Webology, Volume 12, Number 2, December 2015.

[4] A.Pak and P. Paroubek., "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010.

[5] Liza Wikarsa and Sherly Novianti Thahir, "A text mining application of emotion classifications of Twitter's users using Naïve Bayes method", 1st International Conference on Wireless and Telematics, Ieee,Nov 2015.

[6] Zhao Jianqiang, "Pre-processing Boosting Twitter Sentiment Analysis", International Conference on Smart City/SocialCom/SustainCom, IEEE, Des 2015.

[7] Liza Wikarsa, S. T. Indra and Rinaldo Turang, "Using logistic regression method to classify tweets into the selected topics", International Conference on Advanced Computer Science and Information Systems , IEEE, march 2016.

[8] TajinderSingh and MadhuKumari, "Role of Text Pre-processing in Twitter Sentiment Analysis", Procedia Computer Science, 2016.

[9] Itisha Gupta and Nisheeth Joshi, "Tweet normalization: A knowledge based approach", IEEE, International Conference on Infocom Technologies and Unmanned Systems, Dubai, United Arab Emirates, 2017.

[10] Hussein K. Al-Khafaji, Areej Tarief Habeeb,"Efficient Algorithms for Preprocessing and Stemming of Tweets in a Sentiment Analysis System ", IOSR Journal of Computer Engineering (IOSR-JCE),2017.