Published in the USA
International Journal of Media and Information Literacy
Issued since 2005
E-ISSN 2500-106X
2022. 7(1): 60-70

DOI: 10.13187/ijmil.2022.1.60
https://ijmil.cherkasgu.press

Enhancing Information Preservation in Social Media Text Analytics Using Advanced and Robust Pre-processing Techniques

Shah M. Emaduddin [a, *], Rafi Ullah [b], Ibtesam Mazahir [c], Muhammad Zain Uddin [d]

[a] College of Computing and Information Sciences, Karachi Institute of Economics and Technology, Karachi, Pakistan
[b] Sr. Data Scientist at Optimizia, Karachi, Pakistan
[c] Bahria University, Karachi, Pakistan
[d] Institute of Business Administration, Karachi, Pakistan

Abstract

Data mining has become an essential element of today's information world. Different industries and sources daily produce a huge amount of data. When it comes to textual analysis, internet users produce a large amount of data in the form of Twitter Tweets, updates, posts, and comments from Facebook and blogs, short messages, and emails. Analysis of such data will give more valuable information and insights about the studied subject but the problem with social media text is that it is availbel in very raw form. Social media users usually do not produce text in a particular format required by analytics algorithms. Social Media text contains usually miss-spelt words, links, and hash-tags, mentioning people, word/phrase short forms, word elongations, emotional symbols, and many other raw forms. When available text pre-processing techniques (tokenization, lower case, stemming, lemmatization, stop word removals, and normalization) are applied to this raw and un-cleaned data, the removal of many words/phrases results in information loss or information modification. Hence, the curse of data dimensionality vanished and make it difficult to get as much as possible insights from data. We have proposed some advance and robust pre-processing techniques used to increase information preservation from social media text while preserving the semantics of data remain the same.

Keywords: preprocessing, text analysis, natural language processing, sentiment analysis, social media text analysis.

1. Introduction

Preprocessing is a technique in which raw data (improper data) is converted into a proper and structured data form. Machine-Learning based algorithms can better be applied when the data is in proper form to get improved insights. Data used in Natural Language Processing should be converted to lower case alphabets (Angiani et al., 2016; Hadi et al., 2017; Kadhim, Ismael, 2018), normalized (Desai et al., 2015), stemmed (Rani et al., 2015), lemmatized (Angiani et al., 2016; Hadi et al., 2017; Kadhim, Ismael, 2018), and free from all stop words (Rauth, Pal, 2017; Sharma et al., 2015). Preprocessing is mostly applied as a prior step before applying any machine-learning algorithm (Allahyari et al., 2017; Brahimi et al., 2016; Sundari et al., 2017). Textual Analysis is mostly applied on social media text, as of today's world web 4.0 is the era of social interaction and people around the globe communicate with each other via social platforms like Twitter, Facebook and Online forums etc. This results in a huge amount of textual data in the form of comments,

* Corresponding author
E-mail addresses: shahmuhammademad@gmail.com (S.M. Emad)

short messages, emails and discussions on different topics. This data can be brought into valuable form by applying text analytics techniques (Gentzkow, et al. 2017; Batrinca et al., 2015). Textual Analytics domain includes sentiment analysis (Dashtipour et al., 2016; Dickinson et al., 2015; Kharde et al., 2016; Vaghela, et al., 2016), Opinion Mining (García et al., 2016; Lucas et al., 2015; Varathan et al., 2017), Text Summarization (Shetty, Bajaj 2015), Search Engines (*Google, Bing* and others) (Haveliwala, 2003), emotion detection and many other domains (Tabbasum et al., 2019) . Problems often observed in text analysis while dealing in social media text is that users are usually very casual in writing styles. Such as they use emoticons for expressing feelings like (J for I am happy), (L for feeling sad), symbols for expressing states, word/phrase short forms such as (OMG for Oh My God), (GWS for Get Well Soon), Word elongation as the expression of strong feeling such as (I am soooo happyyyyyyy), images, links, hashtags (which sometimes shows some meaning information). User mostly doesn't care about the spelling of words. It means that social media is producing a very huge quantity of data but unfortunately in very raw and un-cleaned form. Hence, whenever it comes to textual analytics, knowledge workers mostly apply state-of-the-art natural language processing pre-processing techniques discussed above. By applying these techniques most of the data is removed, which results in information loss or information modification. For example user of social websites, blogging sites, social review sites allows everyone to express their feelings and thoughts to the world, hence anything removed from their text may change complete context, which leads to wrong analytics. Hence, the huge amount of data gives us a small number of insights/information by using a huge amount of resources in terms of computation and memory. The question arises that why should we not take advantage of alteration of data by keeping the semantics of data in original form. This paper proposed some novel preprocessing steps that can sufficiently increase the information gain of text and minimize the risk factor of wrong analysis, which results in better performance in all textual analytics domains such as emotion detection, sentiment analysis, opinion mining, text summarization, and search engines.

The rest of the paper is organized as follows. Section 2 discuss the literature review of available pre-processing steps in NLP. Section 3 discusses pre-processing steps and their effects on textual data and problems when applied to social media text. Section 4 discusses the proposed pre-processing steps and details. Section 5 demonstrate an evaluation of each proposed step and their contribution to overall information gain and Section 6 is concluding this paper and also puts light on future work.

2. Material and methods

This paper suggests advance and robust pre-processing steps which not only increase information gain of text but also keep the semantic same as it was intentionally published on social media. We have proposed a special ordered sequence of these steps. Pre-processing valuable for social media text analytics includes the following. Adding Strong Feelings. It has been observed that social media users usually used multiple signs such as spaces, exclamation marks (!!!!!), question marks (?????) for adding more strong feelings to text. For example, multiple question marks show the intensity of curiousness about the response (shows asking with excitement). Similarly, multiple exclamation marks show the intensity of excitement. Simple Regular expression can help us to replace these multiple entries with the single symbol (we replace it using a single symbol so that other steps add more value to text later on)

re.sub(SYMBOL+,SYMBOL,text)

Where "re" is regular expression library used in Python and SYMBOL can any letter (occurs multiple times and you want to replace it).

SYMBOL = {SPACE, TAB, ?, !, ., ", ... }

+ in SYMBOL+ shows one or more occurrences of given symbol.

Sharing weblinks on social media in comments, tweets and text as a reference is a common practice of social media users. Very little information can be extracted from links, so it is better to remove all these links. Regular Expression is used to remove these links.

A. Remove @ mentions

On social media, users mention each other to communicate with using mention symbol (@), it has no information at all except it shows the person user name which can be any combination of letters, digits etc. Removing these tokens increase information gain of text, hence reduces complexities.

re.sub('@[^\s] +,'',text)

B. Hash tags replacement and removal

Hashtags also frequently used by social media users to make easy searches. Sometimes these hashtags can be cleaned (remove hash in front of the word) and sometimes it is necessary to remove these words. All the tokens are checked for hashtags removal after hashtags removed token is checked against English dictionary if it is a valid word, it is kept in the text otherwise we discard that tokens. Example hashtags are #style, #instagood, #like4like, #photooftheday.

re.sub(r^'#([^\s] +),r'\1',hashTagToken)

C. Short Forms Expansion

Users on social media also use a short form instead of completed word or phrases to convey message (Desai et al., 2015), hence destroying the analysis power of text. This is not only used in social media but also in literate. Get rid of stuff like "what's" and making it "what is', OMG stands for Oh My God, ILY is for I love you, we use a custom routine that replaces token by it full forms by maintaining a dictionary of contraction patterns and their full form. Dictionary makes it possible to search in constant time. A portion of the dictionary is given below in Table 1.

Table 1. Dictionary of Short Hand Notations

Contraction pattern	*Full Form*
Can't	Cannot
'll	Will
Ain't	Is not
ILY	I love you
&	And
've	Have
Won't	Will not
ROFL	Rolling on floor laughing
NVM	Never Mind
OFC	Of Course
LMK	Let me know
GWS	Get well soon

D. Emoji's Replacement

Emoticons are frequently used for expressing your feelings in comments, tweets, responses etc. The importance of emoji's cannot be ignored while working on text analytics. Each emoji carries its own meaning and importance, it should be considered for analysis. Most of the available libraries consider these such as (Kulkarni, Shivananda, 2019) in sentiment analysis and in (Gelbukh, 2006). In this paper, we replace emoji by it meaning (word or phrase). Custom dictionary and custom routine is used to replace emoji's by the actual meaning of emoji. A sample of the dictionary has been shown in Table 2.

Table 2. Dictionary of Emoji and semantics

Emoji Symbol	*Meaning (Phrase / word)*
☺ or :-)	Basic smile
<3	Love
{y}	Like (Facebook syntax)
:->	Sarcastic
:-#	My lips are scaled
(:-(	Very unhappy or sad
,-)	Winking happy
-)	Winking Smile
:-O	Talkative

E. Elongated words to original words

Users also use elongations to show their excitement towards a post (such as niceeeeee, soooo cuteeee etc). This shows strong feeling in text, although these are not proper English words.

In most of the cases these words are just ignored which results information loss and data size reduction. Instead of removing these words, conversion to base words is more efficient. There are certain Rules decided for word elongation conversion. We use NLTK wordnet corpus for this and Regular Expression (Loper, Bird, 2002). Some of the examples are shown in Table 3 that are converted into base words using proposed pre-processing rule based step.

Table 3. Elongated text to normal text

Elongated Word	*Base Word*
i am sooo sorryyyyy	i am so sorry
veryyyyyy cuuuttteeee	very cute
feeling crazyyyyyyyyyyyyyyy	Feeling crazy
Hhhhhhhooooooootttttt	Hot

F. Spell Correction

Since misspelled words, not in vocabulary can be found frequently in a social media text. This text is meaningless until spelled correctly. So we implemented different spell correction techniques that are faster enough although this is a one-time process (efficiency does not matter here, but still we should use fast techniques). We come across different techniques such as Naïve approach, PeterNorvig and Systematic Delete Spelling Correction (SysSpell). We use PeterNorvig spell correction technique for correcting all tokens. Algorithm correct spelling by generating all possible terms with edit distance technique (inserts + deletes + replaces + transposes) from the token and search them in the dictionary. The correctness mainly depends upon the content of this dictionary. For a word of length n, analphabet size a, an edit distance d=1, there will be n deletions, n-1 transpositions, a*n alterations, and a*(n+1) insertions, for a total of 2n+2an+a-1 terms at search time (http://norvig.com/spell-correct.html).

G. Remove stop words

Step followed by spell correction is the stop words removal. Useless data is referred to as stop words. Stop words removal is the removal of these useless data (words) is referred to as stop-words removal. We would not want these words taking up space in our database, or taking up the valuable processing time. For this, we can remove them easily, by storing a list of words that you consider to stop words. NLTK (Natural Language Toolkit) (Eder et al., 2016) in python has a list of stop words stored in 16 different languages.

H. Lemmatization

Lemmatization refers to doing things properly with the use of a vocabulary and morphological analysis of words/tokens, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma. In computational linguistics, lemmatization is the algorithmic process of determining the lemma of a word based on its intended meaning. Unlike stemming, lemmatization depends on correctly identifying the intended part of speech and meaning of a word in a sentence, as well as within the larger context surrounding that sentence, such as neighboring sentences or even an entire document. As a result, developing efficient lemmatization algorithms is an open area of research.

3. Discussion

Authors like Shetty and colleagues (Shetty, Bajaj, 2015) implemented the idea of categorization, preprocessing, feature matrix, fuzzy logic and sentiment analysis on the textual data. Categorization is performed by using term frequency, summarization with fuzzy systems and finally sentiment analysis is done by using SentiWordNet (Esuli, Sebastiani, 2006). Their system did not contain lemmatization, tokenization and normalization processes, as they helped the system more significance.

Preprocessing can also be done using tokenization, stemming and stopwords removal on the textual data. At first, a document is selected for data extraction, after removing stopwords meaningful words are extracted using the Classis Model. Later stemming function and TF-IDF (Term Frequency-Inverse Document Frequency) is applied (Aizawa, 2003). The disadvantage of this system is that it also removed those words that were less meaningful.

These methods have two categories, first is affix removal methods including Porter Stemmer, Lovins Stemmer, Paice/Husk Stemmer and Dawson Stemmer and the last method is a statistical

method that includes HMM Stemmer, N-gram Stemmer and YASS Stemmer (Rani et al., 2015). Their disadvantage is that they did not discuss some of the methods in detail and did not show their working process.

Authors like Singh and Garg (Singh, Garg, 2018) worked on preprocessing techniques for their data. They used Logistic Classifier for their farmer dataset. They took a dataset from the agriculture sector for farmer queries. They used machine learning packages, python tools, and its libraries. Preprocessing techniques they used were tokenization, lower case conversion, removing punctuation, stopword removal, TF-IDF vector. Their system only extracts the stop words and punctuations and shows the words in the given query.

Sharma and colleagues (Sharma et al., 2015) worked on the stemming and stopword removal techniques by taking 64 documents on iPad. They created a document term matrix that contained 9998 features. They set threshold values from 10 to 90 with a difference of 10. A threshold value is a percentage value, not a parity value. Their data was limited.

Some authors (Allahyari et al., 2017), discussed the knowledge related to data mining, techniques and methods that are efficient in data mining. They worked on text representing and encoding, classification, clustering, biomedical ontologies and text mining for biomedicine and health care information extraction of the text. They briefly discussed these techniques and methods.

Tokenization could also be a useful technique to improve the efficiency of text mining (Sawalha, 2017). The authors discussed different tools of tokenization. These tools contain Nlpdotnet Tokenizer, Mila Tokenizer, NLTK Word Tokenize, Text Blob Word Tokenize, MBSP Word Tokenize, Pattern Word Tokenize, and Word Tokenization with Python NLTK and Stylometry (Eder et al., 2016).

Kadhim and Ismael (Kadhim, Ismael, 2018) selected different documents from different categories and divided them into two models testing and training models. Then applied some text mining techniques including tokenization, stop words removal, stemming, and at the end representing each document as a vector. For the extractions of features, they applied two methods chi-square and TF-IDF. They used BBC English Dataset.

Pre-processing used in NLP

State of the art pre-processing steps are used in natural language processing (Lucas, et al. 2015), which played an efficient role in textual analytics. Techniques observed in the literature are the following.

Tokenization

The process of breaking up a text into pieces such as words, phrases, slang, symbol, digits etc. (Vijayarani, Janani, 2016). They are separated by white space, line breaks, or punctuation marks. Tokens can be made from numbers, alphabets; special characters etc. token can be separated by a mathematical operation because a single token works as a separator in many programming languages.

For example sentence, “whatever you are, be a good one.” will be tokenized as {whatever, you, are, be, a, good, one}.

Lower Case Conversion

Text can be found in any case, as there is no formal rule for writing your expression over social media, but we have to maintain text in a single deterministic format. Therefore it is necessary to convert any text into the lower case so it can easily be readable or accessible for the process. The sound or meaning of the text will remain the same.

For example, the sentence “EVEry MinUTe GOOD sURprise” will be converted into “every minute good surprise”.

Stop-word Removal

Words that are filtered before the processing of natural language of any text due to less information gain is termed as stop-words (Rauth, Pal., 2017). These words are used very frequently and add no uniqueness to the problem. For example “is” is frequently used in every communication or document for example in “Physics” and in “Chemistry”, so it cannot differentiate these two classes and the measures such as TF (Term Frequency), IDF (Inverse Document Frequency), TF-IDF (Term Frequency – Inverse Document Frequency) will be relatively high as compared to other tokens. Most of the search engines avoid these words as they have no proper meaning (Aizawa, 2003).

Table 4. Effect of stop words removal process

Sentence Before Stop-words removal	*Sentence After Stop-words removal*
I love reading books	love reading books
He is suffering from fever	suffering fever
They were eating	Eating

Stemming

Stemming reduces inflected words to their root or original words. As there are multiple forms of a single word (used in a different context), but the meaning remains the same, so we may have a large dataset with redundant tokens (Rauth, Pal., 2017).These redundant words/tokens will affect the processing negatively in terms of space and time complexity. Stemming helps in reducing this extra burden of the process. It extracts information from large datasets and very helpful for the retrieval of the original text. When a word is converted into its original form, its real meaning is useful for data mining else, it will be of no use. Stemming is actually a rule-based process. A large number of stemmer have been observed during the literature of this research, some of the important stemmers are Porter Stemmer, Lovins Stemmer, Paice/Husk Stemmer and Dawson Stemmer, HMMStemmer, N-gram Stemmer and YASS Stemmer.

Table 5. Stemming Rules

Form	*Suffix*	*Stem*
Studies	-es	Studi
Studying	-ing	Study
Fixed	-ed	Fix

Lemmatization

It is the process of grouping the inflected words to their original form or root form. Lemmatization depends on the correct independent part of the speech and the meaning of a word in a sentence.

Table 6. Words to it's base forms

Word	*Lemma*
Help	help(v)
Helps	help(v)
Helping	help(v)
Helped	help(v)

Normalization

Raw data is hard to proceed with different queries, so normalization is a must to get the correct data. Textual data may contain spell errors, abbreviations, or may contain incorrect syntax that further needs to be processed in order to achieve better information gain. For this purpose, we use normalization, which is a process in which text is transformed into its proper and actual form.

Table 7. Text Normalization Process

Raw data	*Normalized form*
Ei8	Eight
Ni8	Night
Gud	Good
☺	Smile

When these techniques are applied to social media text (comments), they may not work properly. For example when English stop removal is applied on the following tokens:

['i', 'ammm', 'sooo', 'happy']

It will results,

['ammm', 'sooo', 'happy']

Here *"ammm"* and *"sooo"* are not removed in fact they are stop-words. These type of words are written differently by different users, hence it will increase the volume of data (also increase memory and computation requirements).

Porter stemmer (Porter, 1980) has no effect when applied on comment from twitter *{Neeeeeeeeeed a diet plannnnnnnn}*. Similarly when applied on *{lovinggggggg and cariiiiing}* stemmed to *{lovinggggggg and cariiii}*, ideally it should stemmed to *{love and care}*.

When NLTK word NetLemmatizer is applied on following words from social media, it fails to do lemmatization properly (Loper, Bird, 2002).

Table 8. Effect of lemmatization on elongated words

Word Before Lemmatization	*Word After Lemmatization*
Happyyyyyyy	happyyyyyyy
Cuteeee	cuteeee
Shifting	shifting

Given is one of the comments from Facebook *{OFC I've decided I'm going to collage sighnin up tomarow},* when this is processed normally as discussed above,

After lower case conversion

{ofc i've decided I'm going to collage sighnin up tomarow},

After applying tokenization,

{ofc, 'i', '", 've', 'decided', 'i', '", 'm', 'going', 'to', 'collage', 'sighnin', 'up', 'tomarow'}

Tokens such as "ve", "", "m" become meaningless, which will be removed later on or will be count as valid tokens. When stop-words are removed,

{'ofc', 'i', '", 'decided', 'i', '", 'going', 'collage', 'sighnin', 'tomarow'}, lemmatization results same set of tokens.

Table 9. Examples comments from social media and source

Comment/Tweet/Text	*Source*	*Observations*
Yep translate option removed by Goggle. Help wanted ad for Goggle "Only Stupid People Need Apply". I don't know why thet keep F'ing things up.	Random Video from Youtube	F'ing is the short form used for filling. Goggle is misspelled in this context.
Why did they remove the translation feature???!!! It was such a helpful feature to have on an international platform like YouTube	Same Video Comment as above	Multi-exclamation marks and question marks.
Such a great day :) and look how far we've fallen from then: D OBAMA!!!!	Video titled as "Obama First speech as president"	Emoticons symbols, which has information but will be removed.
Congratulations ☺ ☺ Modiji and BJP candidates and all supporters.... It's shore next MLA and MP election we are not going to vote just by seeing face of Modiji.... They have to work hard for their areas. Solve the problems of people... Must and should..	Video titled as "Modi First speech as prime minister"	Symbol used for clapping, good luck, gift/roses which carry information but will be removed during traditional preprocessing, Abbreviations also observed such as MLA and MP.
Gud luck to evry 1 getting ther resultz 2morrow	Randomly searched twitter tweet	Misspelling, use digits instead of partial word that sounds similar to original word but different writing style. People on social media mostly

		used this type of words.
I've decided I'm going to collage sighnin up tomarow	Randomly searched twitter tweet	Words Short forms and misspelled words
Welcome Mr. Imran Khan. U r the real inspiration. Pakistan needs u desperately. I salute U from Kashmir!!!!	Imran khan first speech as prime minister	Multi-exclamation marks which shows excitement which should not be ignore in many cases such as emotion detection.
1st time heard the real vision of public servant what they should do.. Hope you accomplish all the promises..	Imran khan first speech as prime minister	Digits and symbol (used for like and good luck)
I cutted my hair and then went back curley	Twitter Tweet	Grammatical mistake (cutted)
Neeeeeeeeeed a diet plannnnnnnn #smart	Random Facebook post	Elongated words, which shows strong feeling in a sentence (cannot be ignored)
Go to dentist twomaro	Twitter retweet	Misspelled word but sounds same as original word
I can think of what desine I want #LifeStyle	Random Facebook comment	
ILY	Reply to a Facebook Comment	Short form of I Love You

4. Results

Proposed pre-processing steps flow has been applied on individual data results are then compared, as claimed the proposed technique out performed in terms of information gain and corpus size. Just for evaluation purpose, an imaginary paragraph is created using given comments and messages from social media and then compared results.

Traditional Pre-processing steps:

Lower case - Tokenization – Stemming – Lemmatization

Proposed technique steps:

As given in methodology

Table 1. Comparative Analysis of Traditional and proposed pre-processing

Input	*Traditional Pre-processing*	*Proposed Technique*
OFC I've decided I'm going to collage tomarow :) https:google.com	ofc decid go collag tomarow :) http : google.com	cours decid go colleg tomorrow basic smile
Neeeeeeeeeed a diet plannnnnnnn #smart	neeeeeeeeeed diet plannnnnnnn # smart	need diet plan smart
Welcome Mr. Imran Khan. U r the real inspiration. Pakistan needs u desperately. I salute U from Kashmir!!!!	welcom mr. imran khan . u r real inspir . pakistan need u desper . I salut u kashmir ! ! ! !	welcom mr. imran khan . real inspir . pakistan need desper . salut kashmir excit
ILY <3	ili < 3	love love
Gud luck to evry 1 getting ther resultz 2morrow	gud luck evri 1 get ther resultz 2morrow nvm	good luck everi one get thier result tomorrow never mind
Such a great day :) and look how far we've fallen from then :D OBAMA!!!!	such great day :) look far 've fallen : D obama ! ! ! !	great day happi look far fallen smile obama excit

Neeed neeeeeed neeeed loveeeee love loveee loveeeee loveeeeee loooovvvvveeee 👆	neeed neeeeeed neeeed loveeeee love lovee loveeee loveeeee loooovvvvveee	need need need need love love love love love good luck
ROFL	Rofl	roll floor laugh

Corpus size after traditional pre-processing steps (Considering unique tokens (Bag of Words) (Huang, Lee, 2008)

Ofc, decid, go, collag, tomarow, :,), http, :, google.com, neeeeeeeeeed, diet, plannnnnnnn, #, smart, welcome, mr., Imran, khan, ., u, r, real, inspire, ., Pakistan, need, u, desper, ., I, salut, u, Kashmir, !, !, !, !, ili, <, 3, gud, luck, evri, 1, get, ther, resultz, 2morrow, nvm, such, great, day, :,), look, far, 've, fallen, :, D, obama, !, !, !, !, neeed, neeeeeed, neeeed,, loveeee, love, lovee, loveeee, loveeeee, loooovvvvveee, rofl

Corpus size is 76 unique tokens and contains amphioxus tokens as well, if we use this corpus for any problem such as language modeling sentiment analysis etc. It may not give us more information. If we ignore not in the vocabulary words, data size will greatly reduce, means that most of the data will be discarded.

Corpus size after proposed pre-processing steps (Considering unique tokens (Bag of Words) (Huang, Lee, 2008).

Cours, decid, go, colleg, tomorrow, basic, smile, need, diet, plan, smart, welcome, mr., Imran, khan, real, inspire, Pakistan, desper, salut, Kashmir, excit, good, luck, everi, one, get, their, result, never, mind, love, great, day, happi, look, far, fallen, Obama, roll, floor, laugh

Corpus size reduced from 76 to 43 which is almost 50% reduction and increasing information gain on overall text. Text contain more information hence more valuable insights can be drawn from.

5. Conclusion

Pre-processing is the important step in any of the machine learning algorithm and analysis technique, hence it must be carried out in a fashion so that the overall integrity of data remains the same and make data understandable by the machine. Several preprocessing steps are used by researchers to improve data quality. In the era of social media where people use shorthand notations, misspelled, words that are not in the vocabulary, symbols and multiple languages in a single post/message/comment, it becomes difficult for researchers to apply these traditional pre-processing steps. By applying these steps, most of the data (valuable data) will be removed. Hence, data size reduced which will definitely decrease textual analytics. This paper proposed some novel pre-processing steps and an efficient ordered sequence of these steps to make use of that noisy data / unclean data. By applying these techniques, we make capable existing algorithms to derive more insights from social media data. The proposed technique is shown efficient in terms of accuracy and information gain while performing textual analytics. In the future, we will perform different text analytics domain such as language classification, intent detection, emotion classification, depression detection etc. on raw data (data result after traditional pre-processing steps) and mature data (data after proposed pre-processing steps) and will compare different algorithms in term of accuracy, analytical information and information gain from these texts. One of the problems still exists that in social media text contain text in different languages by participants while discussing even single topic, in future novel normalization step will be incorporated in this technique to normalize complete text in term of language (language detection will be performed and all the text will be converted to single common patter for textual analysis).

References

Aizawa, 2003 – *Aizawa, A.* (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*. 39(1): 45-65.

Allahyari et al., 2017 – *Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.* (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv preprint ArXiv*.

Angiani et al., 2016 – *Angiani, G., Ferrari, L., Fontanini, T., Fornacciari.* (2016). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. *KDWeb*.

Batrinca, Treleaven, 2015 – *Batrinca, B., Treleaven, P.C.* (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & Society*. 30(1): 89-116.

Brahimi, et al. 2016 – *Brahimi, B., Touahria, M., Tari, A.* (2016). Data and text mining techniques for classifying Arabic tweet polarity. *Journal of Digital Information Management*. 14(1).

Dashtipour, et al., 2016 – *Dashtipour, K., Poria, S., Hussain, M., Cambria, M., Hawalah, A., Gelbukh, A., and Zhou*, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*. 8(4): 757-771.

Desai, Narvekar, 2015 – *Desai, N., Narvekar, M.* (2015). Normalization of noisy text data. *Procedia Computer Scienceю* 45: 127-132.

Dickinson, Hu, 2015 – *Dickinson, B., Hu, W.* (2015). Sentiment analysis of investor opinions on twitter.S *ocial Networking*. 4(3): 62.

Eder et al., 2016 – *Eder, M., Rybicki, J., Kestemont, M.* (2016). Stylometry with R: a package for computational text Analysis. *R Journal*. 8(1).

Esuli, Sebastiani, 2006 – *Esuli, A., Sebastiani, F.* (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *In LREC*. 4: 417-422.

García et al., 2016 – *García, S., Ramírez-Gallego, S., Luengo, J., Mannuel, B., Francisco, H.* (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*.

Gelbukh, 2006 – *Gelbukh, A.* (2006). Computational linguistics and intelligent text processing. *7th International Conference Proceedings*. 3878: 19-25. Springer.

Gentzkow et al., 2017 – *Gentzkow, M., Bryan, T.K., Matt, T.* (2017). Text as data. w23276. *National Bureau of Economic Research*.

Hadi et al., 2017 – *Hadi, R.M., Hashem, S.H., Maolood, A.T.* (2017). An effective preprocessing step algorithm in text mining application. *Engineering and Technology Journal,* 35(2): 126-131.

Haveliwala, 2003 – *Haveliwala, T.H.* (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*. 15(4): 784-796.

Huang, Lee, 2008 – *Huang, C., Lee, L.H.* (2008). Contrastive approach towards text source classification based on top-bag-of-word similarity. *In Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*.

Kadhim, 2018 – *Kadhim, A.I.* (2018). An Evaluation of preprocessing techniques for text classification. I*nternational Journal of Computer Science and Information Security*. 16(6).

Kharde, Sonawane, 2016 – *Kharde, V., Sonawane, S.* (2016). Sentiment analysis of twitter data: A survey of techniques. *ArXiv preprint*.

Khedr, Yaseen, 2017 – *Khedr, A.E., Yaseen, N.* (2017). Predicting stock market behavior using data mining technique and news sentiment analysis. *International Journal of Intelligent Systems and Applications*. 9(7): 22.

Kulkarni, Shivananda, 2019 – *Kulkarni, A., Shivananda, A.* (2019). Advanced Natural Language Processing. *Natural Language Processing Recipes*. APress, Berkeley: 97-128.

Loper, Bird, 2002 – *Loper, E., Bird, S.* (2002). NLTK: the natural language toolkit. ArXiv preprint cs/0205028.

Lucas et al., 2015 – *Lucas, C., Nielsen, R.A., Roberts, M.E., Stewart, B.M., Storer, A., Tingle, D.* (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*. 23(2): 254-277.

Porter, 1980 – *Porter, M.F.* (1980). An algorithm for suffix stripping. *Program*. 14(3): 130-137.

Rani et al., 2015 – *Rani, S.B., Ramesh, M., Sathiaseelan, J.G.R.* (2015). evaluation of stemming techniques for text classification. *International Journal of computer science and mobile Computing*. 4(3): 165-171.

Rauth, Pal, 2017 – *Rauth, A.P., Pal.*, A. (2017). Design and evaluation of text pre-processor: a tool for text pre-processing. *AMSE Journals IIETA Publication*. 54(2): 169-184.

Sawalha, 2017 – *Sawalha, N., Sawalha, M.* (2017). A Study of Arabic Keyboard. *New Trends in Information Technology (NTIT)*.

Sharma, Jain, 2015 – *Sharma, D., Suresh, J.* (2015). Evaluation of stemming and stop word techniques on text classification problem. *International Journal of Scientific Research in Computer Science and Engineering*. 3(2): 1-4.

Shetty, Bajaj, 2015 – *Shetty, A., Bajaj, R.* (2015). Auto text summarization with categorization and sentiment analysis. *International Journal of Computer Applications*. 130(7): 57-60.

Singh, Garg, 2018 – *Singh, Y., Garg, N.K.* (2018). Preprocessing farmer query data using classic method and building classifier model. *IJSRCSEIT*. 3(3).

Sundari, Guna, Sundar, 2017 – *Sundari, D., Guna, J., Sundar, D.* (2017). A study of various text mining technique. *International Journal of Advanced Networking & Applications (IJANA)*. 8: 82-85.

Tabbasum et al., 2019 – *Tabbasum, H., Emaduddin, S.M., Awan, A., Ullah, R.* (2019). Multi-Class Emotion Detection (MCED) using Textual Analysis. *International Conference on Computing and Information Science.*

Vaghela et al., 2016 – *Vaghela, V.B., Jadav, B.M, Scholar, M.E.* (2016). Analysis of various sentiment classification techniques. *International journal of Computer applications*. 140(3).

Varathan et al., 2017 – *Varathan, K.D., Giachanou, A., Crestani, F.* (2017). Comparative opinion mining: a review. *Journal of the Association for Information Science and Technology*. 68(4): 811-829.

Vijayarani, Janani., 2016 – *Vijayarani, S., Janani, R.* (2016). Text mining: open source tokenization tools-an analysis. *Advanced Computational Intelligence An International Journal*. 3(1): 37-47.