Improved Approach for Preprocessing Data

Mr. SanketPatil (*Author*)
Computer Engineering,
SLRTCE,
Mumbai, India

Prof. VarshaWangikar
Information Technology,
K.C.College of Engineering and
Management Studies and Research
College,
Mumbai, India

Prof. K.Jayamalini
Computer Engineering,
SLRTCE,
Mumbai, India

Abstract - "No quality data, no quality results". Preprocessing is most important task in Tweet data Analysis as it passes data to further stages. It focuses on transforming data into a form which can be easily & effectively used as input in many domains. In this paper, we have introduced novel algorithm for the preprocessing of the text using distributed approach. This algorithm makes use of many techniques to remove redundant information used for Tweet data analysis and it will result in compressing a lengthy statement into shorter and easier one without changing its meaning. Filtered sentences enhance the performance of sentiment analysis system in terms of time & space complexity also. Further this type of data can be used for different types of research domains like natural language processing, information retrieval, text classification and text clustering

Keywords*— SA- Sentiment Analysis, NLTK- Natural Language Processing Tool, Twitter*

*****

I. Introduction

Huge amount of reviews are available on Internet which can be used to determine overall opinion or feeling of product. Very often these reviews are raw so to preprocess them is one of the challenging tasks. We need to analyze them so that they can be used in intelligent decision making system. The main objective of preprocessing is to obtain the key features or key terms from existing text documents and to enhance the relevancy between word and its associated class. Pre-Processing step is crucial in determining quality of output for next stage that is classification. It's important to select specific words which convey meaning & remove other words which do not contribute to distinguish text between documents. The pre-processing phase thus converts the original data in a text-mining ready structure. Process of text mining is depicted in the following figure 1[1]. It includes gathering of data from different sources, preprocessing text, applying text mining techniques to learn meaning of sentence. This data is then analyzed to gain specific expected information.

The importance of preprocessing is emphasized by the fact that the quantity of training data on web is growing exponentially with the dimension of the input space. It has already been proven that the time spent on preprocessing can take from 50% up to 80% of the entire classification process [2], which clearly evidences the importance of preprocessing in text classification process.

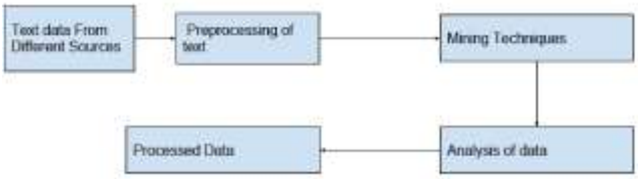

Fig. 1 Basic Mining Process

This paper discusses the various preprocessing techniques used in the present research work.

1.1 Applications of Text Mining

Information Retrieval

Information retrieval (IR) concept has been developed in relation with database systems for many years. Information retrieval is the association and retrieval of information from a large number of text-based documents. The information retrieval and database systems, each handle various kinds of data; some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in conventional database systems, such as unstructured documents, estimated search based on keywords, and the concept of relevance. Due to the huge quantity of text information, information retrieval has found many applications. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines [3].

Information Extraction

The information extraction method identifies key words and relationships within the text. It does this by looking for predefined sequences in the text, a process called pattern matching. The software infers the relationships between all the identified places, people, and time to give the user with meaningful information. This technology is very useful when dealing with large volumes of text. The information being "mined" is already in the form of a relational database is being assumed by Traditional data mining . Unfortunately, for many applications, electronic information is only available in the form of free natural language documents rather than structured databases [5].

Categorization

Identifying the main themes of a document by inserting the document into a pre-defined set of topics is involved in Categorization. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not try to process the actual information as information extraction does. Rather, the categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a glossary for which topics are predefined, and relationships are identified by looking for large terms, narrower terms, synonyms, and related terms [6].

Natural Language Processing

An area of research and application that explores how computers can be used to understand and manipulate natural language text is known as Natural Language Processing (NLP). NLP researchers aim to collect knowledge on how human beings understand and use language so that fitting tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the preferred tasks [6].

The basics of NLP lie in a number of disciplines, viz. computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, etc. Applications of NLP include a number of fields of studies, such as machine translation, natural language text processing and summarization, user interfaces, multilingual and cross language information retrieval (CLIR), speech recognition, artificial intelligence and expert systems and so on[8].

II. LITERATURE SURVEY

Purpose of stemming is to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb etc Anjali Ganesh Jivani [4] has discussed different methods of stemming and their comparisons in terms of use, advantages & disadvantages. She has also focused on difference between lemmatization & stemming.

Vishal Gupta et.al [5] has evaluated the stemmer‟s performance in applications such as spell checker for multiple languages. This approach uses algorithm to remove suffixes using a list of frequent suffixes, He has also discussed common stemming techniques and existing stemmers for Indian languages.

S.Jusoh and H.M. Alfawareh [6] have reviewed the literature on NLP. Natural Language Processing (NLP) is a way of analyzing texts by computerized means. NLP involves gathering of knowledge on how human beings understand and use language. This is done in order to develop appropriate tools and techniques which could make computer systems understand and manipulate natural languages to perform various desired tasks. It also covers a hint about the history of NLP

K.K. Agbele [7] discussed the technique for developing pervasive computing applications that are flexible and adaptable for users. However, information retrieval (IR) is often defined in terms of location and delivery of documents to a user to satisfy their information need, in this context, . In most cases, morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. The algorithm Context-Aware Stemming (CAS) is proposed, which is a modified version of the extensively used Porter's stemmer. Considering only generated meaningful stemming words as the stemmer output, the results show that the modified algorithm significantly reduces the error rate of Porter's algorithm from 76.7% to 6.7% without compromising the efficacy of Porter's algorithm.

Hassan Saif [8] has explored whether removing stop words helps or hampers the effectiveness of Twitter sentiment classification methods. He used six datasets & found that sometimes it gives negative impact on sentiment analysis system as it may happen words conveying meaning will be deleted because of adverse effect of an algorithm.

III. SYSTEM ARCHITECTURE

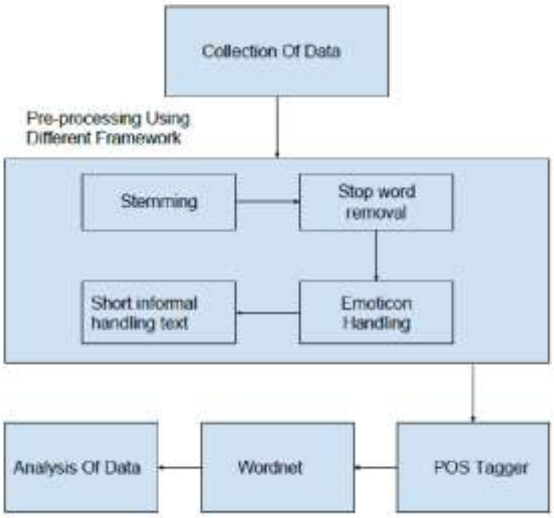

Fig. 2 System Architecture

A. Collection of data:

Input or data sets used in Tweet analysis plays very important role. Main source is coming from reviews from social media sites for example-in political debate, we may find peoples opinion about certain election candidate/political parties [9]. Also, we can predict results of election from posts sent by people. This raw data may contain short text which should be carefully handled and it needs special attention from programmer. Tremendous amount of work has been done in recent years on the problem of text collections in the database and information retrieval communities.

B. Stemming

This method is used to detect the root form of a word. For example, words saw, seen, seeing & see - all can be stemmed to the word "**See**" [10]. The goal here is to remove various suffixes, to reduce the number of words so that to save time & memory. This is demonstrated in Figure 3.

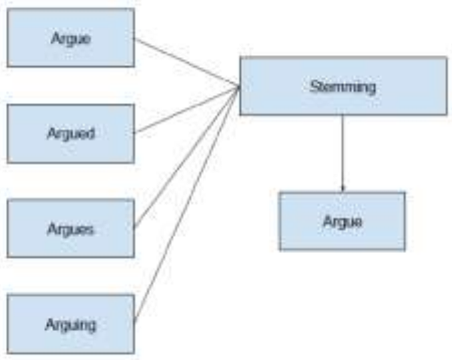

Fig.3 Stemming Example

Here morphological forms of word are treated as semantically related. There are two points are considered while using a stemmer:

1. Words with different meaning are kept separately.

2. Morphological forms of words are supposed to have same meaning so they are mapped to same word/stem.

Stemming is usually considered as a recall-enhancing device. For languages with relatively simple morphology, the power of stemming is less than for those with a very complex morphology. Most of the stemming experiments done so far are for English and other west European languages [11].

Usually, stemming algorithms can be classified into three groups: truncating methods, statistical methods, and mixed methods [12]. Each of these groups has a typical way of finding the stems of the word variants. These methods and the algorithms discussed are shown in the Figure 4.

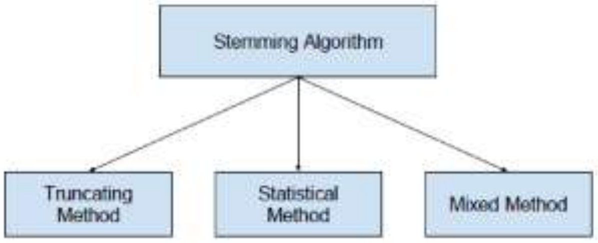

Fig. 4 Stemming Algorithm Methods

- **Truncating Methods (Affix Removal)**

As the name obviously suggests these methods are related to removing the suffixes or prefixes (commonly known as affixes) of a word [12]. The most basic stemmer is the Truncate (n) stemmer which truncated a word at the nth symbol i.e. keep n letters and remove the rest.

- **Statistical Methods**

These are the stemmers who are based on statistical analysis and techniques. Most of the methods remove the affixes, but after implementing some statistical procedure [12].

- **Mixed Methods**

This is another approach in stemming and it involves both the inflectional as well as the derivational morphology analysis. The corpus should be very large to develop these types of stemmers and hence they are part of corpus base stemmers too. In case of inflectional the word variants are related to the language specific syntactic variations like a plural, gender, case, etc., whereas in derivational the word variants are related to the part-of-speech (POS) of a sentence where the word occurs [12].

C. Stop word removal

Most frequently used words in English are useless in Text mining. Such words are called Stop words. Stop words are language specific functional words which carry no specific information. It may be of the types such as pronouns, prepositions, conjunctions etc. If we remove stopwords from our sentence then also it should not affect meaning of sentence. Benefits of removing the stopwords are:-

1) Reduction in size of data.

2) Reduction in processing time.

In fact, it is possible to obtain a compression in the size of the indexing structure. List of 425 stop words is identified by researchers. Despite of these benefit, elimination of stopwords might reduce recall. For instance, consider a user who is looking for documents containing the phrase, "to be or not to be". Elimination of stopwords might leave only the term **be**making it almost impossible to properly recognize the documents which contains the phrase specified [1].

D. Emoticons handling

Nowadays people are becoming bad fans of using emoticons to express their emotions in all types of feedbacks & reviews. But use of emoticons makes processing of data difficult, as system is completely unaware about meaning of these emoticons. In our system we can handle these emoticons by providing specific meaning to every emoticon. Following is the list of some emoticons and their meaning.

Typical examples of emoticon synsets.[15]

Emoticon synset Emoticons

Happiness :-D, =D, xD, (^ ^)

Sadness :-(, =(

Crying :'(, ='(, (; ;)

Boredom - -, -.-, (><)

Love <3, (L)

Embarrassment :-$, =$, >///<

Table 1:Emoticons used in reviews

Positive Emoticons	Negative Emoticons
o n n *-* *O* * *	
:P :D :d :p	:(;(:'(;'(
;P ;D ;d ;p	=(={):);
:-) ;-) :-) ;-)	)':)';)= }=
:<) :>) ;>) =)	;-{{ ;-{ :-{{ :-{
(; :} {: ;}	:,(:'{
{; :]	[: ;]
[; :') ;') :-3	

;-3 :-x ;-x :-X	
;-X :-} ;-} :-]	
;-] :-.)	

E. Short informal text handling.

Short informal text is one of the challenges to sentiment analysis. It is limited in length, usually spanning one sentence or less. It contains any misspellings, slang terms, and shortened forms of words.

Example: How R U?, Gr8

Okkkk, Wowww, Yup! Nops..etc

It is tedious task to process such type of data. So we need to handle it by the same way that we used while handling the emoticons. We can provide specific meaning to different types of short informal texts which user uses while giving his opinion in a review [16]

F. POS Tagger

POS tagging is the task of labeling each word in a sentence with its appropriate parts of- speech like noun, verb, adjective, etc. This process takes an untagged sentence as input then assigns a POS tag to words and produces tagged sentences as output. The most widely used part of speech tagset for English is PennTree bank tagset. The example shown below represents the POS tagging for English sentences [17].

Example *The boy is going to the school.*

Part-of-Speech Tagging

The/**DT** boy/**NN** is/**VBZ** going/**VBG** to/**TO** the/**DT** school/**NN** ./.

G. WordNet:

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.[18]

Table 2:Semantic relations in Wordnet

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

WordNet includes the following semantic relations:

• Synonymy is WordNet's basic relation, because WordNet uses sets of synonyms (synsets) to represent word senses. Synonymy (syn same, onyma name) is a symmetric relation between word forms.

• Antonymy (opposing-name) is also a symmetric semantic relation between word forms, especially important in organizing the meanings of adjectives and adverbs.

• Hyponymy (sub-name) and its inverse, hypernymy (super-name), are transitive relations between synsets. Because there is usually only one hypernym, this semantic relation organizes the meanings of nouns into a hierarchical structure.

• Meronymy (part-name) and its inverse, holonymy (whole-name), are complex semantic relations. WordNet distinguishes component parts, substantive parts, and member parts.

• Troponymy (manner-name) is for verbs what hyponymy is for nouns, although the resulting hierarchies are much shallower.

• Entailment relations between verbs are also coded in WordNet.

H. Analysis of processed data:

Further this data is used for different natural language processing tasks such as classification, clustering, association etc.

IV. Conclusion

Pre-processing activities plays a vital role in the various applications. Therefore it is concluded that the domain specific applications gives more correct results for text mining. This paper presented pre-processing techniques using distributed approach, which significantly reduces time required in preprocessing phase.

References

[1] Nikita P.Katariya et al, International Journal of Computer Science andMobile Applications, Vol.3 Issue. 1, January- 2015, pg. 01-05

[2] Katharina, M. and Martin, S. (2004) The Mining Mart Approach to Knowledge Discovery in Databases, NingZhong and Jiming Liu (editors), Intelligent Technologies for Information Analysis, Springer, Pp. 47-65

[3] Vishal Gupta, Gurpreet Singh Lehal, A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages, Journal of Emerging Techniloigies in Web Intelligence, VOL. 5, NO. 2, MAY 2013.

[4] Anjali Ganesh Jivani , A Comparative Study of Stemming Algorithms, International Journal of Computer, Technology and Application, Volume 2, ISSN:2229-6093.

[5] Vishal Gupta, Gurpreet Singh Lehal, A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages, Journal of Emerging Techniloigies in Web Intelligence, VOL. 5, NO. 2, MAY 2013.

[6] S.Jusoh and H.M. Alfawareh, Natural language interface for online sales, in Proceedings of the International Conference on Intelligent and Advanced System (ICIAS2007). Malaysia: IEEE, November 2007, pp. 224–228.

[7] Agbele, A.O. Adesina, N.A. Azeez, & A.P. Abidoye , Context-Aware Stemming Algorithm for Semantically Related Root Words, African Journal of Computing & ICT.

[8] Hassan Saif, Miriam Fernandez,Yulan He, HarithAlani,OnStopwords,Filtering and Data Sparsity for Sentiment Analysis of Twitter.

[9] Swati B. Bhonde, Prof. J. R. Prasad "Sentiment Analysis - Methods, Applications & Challenges", IJECCE,Vol.6, Issue 6(2015)Pp 634-640.

[10] SalehAlsaleem, Automated Arabic Text Categorization Using SVM and NB,International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011.

[11] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya, Preprocessing Techniques for Text Mining - An Overview

[12] Deepika Sharma, Stemming Algorithms, A Comparative Study and their Analysis, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Foundation of Computer Science FCS, New York, USA, Volume 4– No.3, September 2012 – www.ijais.org.

[13] Harman Donna, How effective is suffixing? Journal of the American Society for Information Science, 1991; 42, 7-15 7.

[14] J. B. Lovins, Development of a stemming algorithm, Mechanical Translation and Computer Linguistic., vol.11, no.1/2, pp. 22-31, 1968.

[15] Text Mining, chapter 4, "Preprocessing for English Sentence"

[16] https://wordnet.princeton.edu/

[17] V.Jude Nirmal1 and D.I. George Amalarethinam, Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data

[18] George A. Miller, ACM, WordNet: A Lexical Database for English