International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 11 Issue 5 ǁ May 2023 ǁ PP. 26-30

A Stemming process in Sentiment Analysis for Social Media Tweets

S. Amutha
Assistant Professor, Department of Computer Science,
Government Arts college for Women, Nilakottai,
Tamilnadu, India

Abstract— *In the Sentiment Analysis processes there are various phases like, Preprocessing, Feature selection, tokenization, lemmatization, stop word removal, POS tagging, removal of URLs, substitution of emoticons, spelling correction, word normalization, abbreviation lookup, punctuation removal especially Stemming processes. In the Sentiment Analysis Stemming is mainly use to trim or find the root of the word. The unwanted letters like es, ing, ed and ful etc. There are different types of stemmers used to find the meaning word in Social Network sites belongs to Sentiment Analysis Mining.*

Keywords— Sentiment Analysis, Stemming, Social Network, classification, opinion mining

---

Date of Submission: 19-04-2023 Date of acceptance: 03-05-2023

I. INTRODUCTION

In this day and age research is faced with both extraordinary opportunities and challenges. The social order is willing to devote in research as the source of a acquaintance economy as long as research proves to be responsive to its needs. One of the most visible trends on the web is the coming out of Social Web sites, which help people, create and gather Knowledge by simplifying user contribution via blogs, Tagging and Folksonomies, wikis, podcasts and Online Social Networks.

Sentiment Analysis (SA) is used to measure the thoughts of people in social media texts in various ways like positive, negative, neutral and real and bias emotions. User can select whether any topic the expression of public will in different views inside the three ways convey in above. Large amount of internet users expressed their views about topics and events which is happening current incidents and accident in sentimental word with exclamatory sentences. Lot of smiley is used to show their lengthy expression in a single or simple of pictures. Most of the multimedia messages give the 100% result of good and bad approach of internet users[1].

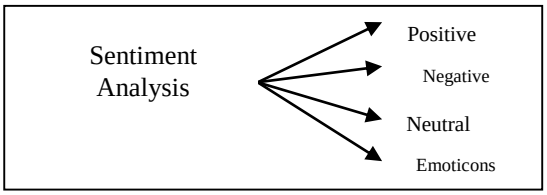

Figure 1. Categories of Sentiment Analysis

II. SOCIAL MEDIA

Twitter is a widely used micro-blogging environment which serves as a medium to share opinions on various events and products. Because of this, analyzing Twitter has the potential to reveal opinions of the general public regarding these topics[2]. However, mining the content of Twitter messages is a challenging task due to a multitude of reasons, such as the shortness of the posted content and the informal and unstructured nature of the language used. The aim of this study is to produce a methodology for analyzing sentiments of selected Twitter messages, better known as Tweets. This project elaborates on two experiments carried out to analyze the sentiment

III. WHAT IS STEMMMING?

Word Stemming is an important feature supported by present day indexing and search systems. Indexing and searching are in turn part of Text Mining applications, Natural Language Processing (NLP) systems and Information Retrieval (IR) systems. The main idea is to improve recall by automatic handling of word endings by reducing the words to their word roots, at the time of indexing and searching. Recall in increased without compromising on the precision of the documents fetched[3]. Stemming is usually done by removing any attached

suffixes and prefixes (affixes) from index terms before the actual assignment of the term to the index. Since the stem of a term represents a broader concept than the original term, the stemming process eventually increases the number of retrieved documents in an IR system. Text clustering, categorization and summarization also require this conversion as part of the pre-processing before actually applying any related algorithm.

IV. WORKING OF A STEMMER

It has been seen that most of the times the morphological variants of words have similar semantic interpretations and can be considered as equivalent for the purpose of IR applications. Since the meaning is same but the word form is different it is necessary to identify each word form with its base form. To do this a variety of stemming algorithms have been developed[4]. Each algorithm attempts to convert the morphological variants of a word like introduction, introducing, introduces etc. to get mapped to the word 'introduce'. Some algorithms may map them to just 'introduc', but that is allowed as long as all of them map to the same word form or more popularly known as the stem form. Thus, the key terms of a query or document are represented by stems rather than by the original words. The idea is to reduce the total number of distinct terms in a document or a query which in turn will reduce the processing time of the final output.

V. STEMMING AND LEMMATIZING

The basic function of both the methods – stemming and lemmatizing is similar. Both of them reduce a word variant to its 'stem' in stemming and 'lemma' in lemmatizing. There is a very subtle difference between both the concepts. In stemming the 'stem' is obtaining after applying a set of rules but without bothering about the part of speech (POS) or the context of the word occurrence[5]. In contrast, lemmatizing deals with obtaining the 'lemma' of a word which involves reducing the word forms to its root form after understanding the POS and the context of the word in the given sentence.

VI. DESCRIPTION OF STEMMING

Stemming is to find the main part of the word gives meaning. It is one of the preprocessing steps in the Text mining which is used in the Sentiment Analysis Mining. It is used in the Information Retrieval systems. The goal of stemming is to reduce the word or trim the word, which in the form of verb, noun, adverb, and adjective etc. It is called root finding. In indexing and searching the most part done by the stemming process the prefixes and suffixes are removing in the action word. It is a broader way to find the original word which is used in the Information Retrieval.

Stemming is a function for collapsing clear-cut word forms. This could be reducing the vocabulary size and thereby sharpening one's results, especially for small data sets. It will be used for emotion mining. It is used to reduce the elapsed time by the reduction the elapsed time by the reduction of dictionary size[7]. It will be used to neutralize the grammatical effect of words in order to get the real and bias emotion. In SA and the complex natural language the feature stemming and pruning also essential to identify similar opinion words and group them together. It reduces the set of opinion words that are used for classification.

For example words like addition, addictive, added, adding, are stemmed to the word add and are considered as a single opinion word. Since all the words have same opinion with same orientation therefore, stemming them will enable the classifier to treat them as the same word attract. It is reducing the features set improve the performance of the classifiers.

1) Stemming will be used for emotion mining.
2) Stemming is used to reduce the elapsed time by the reduction of dictionary size.
3) Stemming will be used to neutralize the grammatical effect of words in order to get the real and bias emotion.

Users can use TACIT's preprocessing features to remove stop words like common words such as "a" or "the" that in most circumstances provide no useful information, perform automatic stemming (i.e) map each word to its root and standardize text to convert all text to lowercase or remove extraneous information. Tacit automatically detects the language of the text and uses an implementation of the porter stemmer or snowball stemming b to apply the correct stemming rules for that language. It is very important to note that this feature is different from LIWC-style stemming because LIWC matches string patterns and does not necessary map words to their roots.

Preprocessing of the source document is done in which linguistic techniques are applied which includes segmentation of sentences, removal of stop words, removal of punctuation marks, stemming etc. Segmentation process deals with dividing the text into sentences. In preprocessing stemming is third part processing made the words into its grammatical root form, it converts word like wonderful-wonder, watch, watcher, watching, watched, watches to root word watch[8].

Stemming is the process of reducing the words with the same root or stem to a common form, thus removing the variable suffixes. Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Stemming have two stages, first it is based upon clustering which exploits the lexical and semantic information of words is used to prepare large scale training data for the second stage algorithm. In second stage uses a maximum entropy classifier[9].

VII. STEMMING ANALYSIS

It is way of reducing the size of the word into its basic form while processing most of its linguistic features by knowing the Word and Pattern is known as Stemming Analysis. It is utility task that is used as a sub system of more complex ones. It is process of reducing large space word into smaller one. It is conceptually similar to hashing mechanism. As the size differs between the two spaces words and roots, this leads to collision effect where two or more words have the same root. However and unlike hashing, collision can positively influence the stemming task.

In the Social Network sites like twitter the tweets connected with any other language to express the thoughts of the user. And it is detected whether it's in spam detection calculate the unigram or bigram in the tokenization part, after that stemming is enter into it for the indexing and query process. This will happen it belongs to Unigram. Same way it is in Bigram or Trigram means there is noise reduction done in the process finally gets the index results.

In Figuration the idioms are typically have figurative meaning stemming from metaphors, hyperboles and otter types of figuration. Stemming will be used for emotion mining, It is used to reduce the elapsed time by the reduction of dictionary size. It will be used to neutralize the grammatical effect of words in order to get the real and bais emotion.

In the MRL (Morphologically rich languages) may result in loss or giving incorrect sentiment meaning to words, user decided to expand the sentiment lexicon.MRL can lead to a potential loss or erroneously assigned of sentiment polarity information because of the large number of inflected and derived words. At end where the affixes of the word are clipped from the word to make it concise with minimum length, yet having the same meaning. Stemming was performed so as to minimize sparsity, the bag of words was constructed with binary frequency and a term is considered frequent if it occurs in more than one tweet[6].

Through tokenization breaks down the document into its basic units, the words are usually in their inflected forms or in a form that describes the tense of the sentence. Stemming is the process of eliminating prefix and suffix from a word to obtain the basic structure of the word, also called the seed word. Seed words are required in order to identify the polarity. This process is carried out in stemming[10].

Stemming reduces inflection in words; however some words are incomplete and ripped after stemming. It is an approach used to reduce a word to its stem or root form and is used widely in information retrieval tasks to increase the recall rate and give us most relevant results. It is the conflation of the variant forms of a word into a single representation. The stem does not need to be a valid word, but it most capture the meaning of the word. It has been extensively used to increase the performance of Information Retrieval Systems.

Lot of world languages like Czech, French, Hebrew, Hungarian, Portuguese and Indian Languages like Bengali, Marathi and Hindi stemming increase the number of documents retrieved by between 10 to 50 time. it is used to reduce the size of index files.

VIII.TYPES OF STEMMER

There are lots of Stemmers supports for stemming process done in the words.

1. Porter Stemmer: it is one of the earliest and best known stemming algorithms. It works by heuristically identifying word suffixes and stripping them off, with some regularization of the endings. Often it collapses sentiment distinctions, by mapping two words with different sentiment into the same stemmed form. when the porter stemmer evolved into a whole stemming framework called SNOWBALL. It allows programmers to develop their own stemmers for other character sets or languages. It is converting the non-root words to root word. The root word which is not carries emotion emphasis mining especially.

Sentiment Analysis, on the data collected from the Social network or Social media like Face book, twitter, pinterest or IMDB sites. It produces the results with accuracy measured in terms of precision and recall. It minimizes the feature set and makes efficient classification performances by using Java language.

2. Lovins Stemmer: spacing to squeeze more text into a limited number of pages. First published stemming algorithm ever is it. Which was designed especially for English stemming, Its needs only two steps for stemming a word according to predefined endings and transformation rules. This makes the algorithm very simple and very fast. It is popular and effective stemmer. It removes the longest suffix from a word. It always removes a maximum

of one suffix from a word, due to its nature as single pass algorithmic is very fast removing duple letters in words like submitting being changed into submit and also holds many irregular plurals ex; children-child, tooth-teeth, louse-lice and goose-geese etc.

3. Paice /Hust Stemmers: It is an iterative algorithm with one table containing about 120 rules indexed by the last letter of a suffix form and every iteration taking care of both deletion and replacement as per the rule applied. It is heavy algorithm and over stemming may occur.

4. Dawson Stemmer: It is extension of Lovins approach. It covers more suffixes than Lovins and its done fast execution. It is very complex and lacks a standard reusable.

5. N-Gram Stemmer: it is a string of n, usually adjacent, characters extracted from a section of continuous text. It is to be precise an n-gram is a set n consecutive characters extracted from a word.

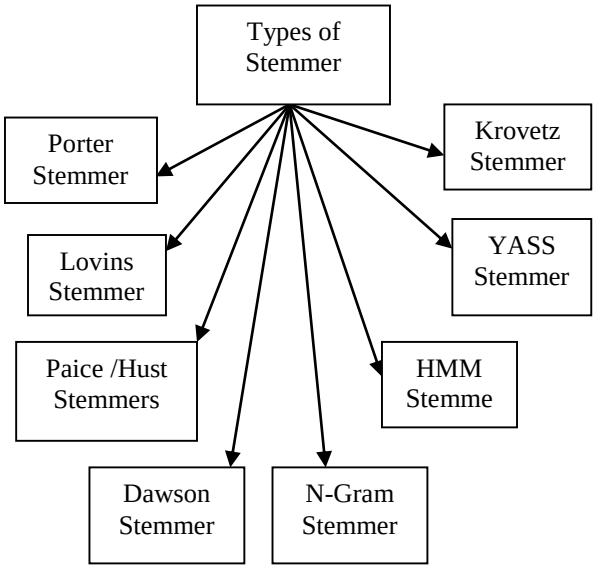

Figure 2. Types of Stemmer

6. HMM Stemmer: Hidden Markov Model, Which are finite-state automata where transitions between states are ruled by probability functions. A sequence of litters that forms a word can be considered the result of a concatenation of two subsequences, a prefix and a suffix.. A way to model this process is through an HMM where the states are divided in two disjoint sets: initial states belong only to the stem set word always starts with a stem. Transitions from states of the suffix set to states of the stem set always have a null probability word can be only a concatenation of a stem and a suffix.

7. Yass Stemmer: Yet Another Suffix Striper. This Stemmer comes under the category of statistical as well as corpus based. The clusters are created using hierarchical approach and distance measures. Then the resulting clusters are considered as equivalence classes and theirs centroids as the stems.

8. Krovetz Stemmer: It is a linguistic lexical validation stemmer. It is based on the inflectional property of words and the language syntax; it is very complicated in nature. It effectively and accurately removes inflectional suffixes in three steps.

1. Transforming the plurals of a word to its singular form.
2. Converting the past tense of a word to its present form.

IX. TRUNCATING METHODS

As the name clearly suggests these methods are related to removing the suffixes or prefixes (commonly known as affixes) of a word. The most basic stemmer was the Truncate (n) stemmer which truncated a word at the nth symbol i.e. keep n letters and remove the rest. In this method words shorter than n are kept as it is. The chance of over stemming increases when the word length is small. Another simple approach was the S-stemmer – an algorithm conflating singular and plural forms of English nouns. This algorithm was proposed by Donna Harman. The algorithm has rules to remove suffixes in plurals so as to convert them to the singular forms

X. 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.

XI. INFLECTIONAL AND DERIVATIONAL METHODS

This is another approach to 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 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.

XII. PROBLEM FORMULATION

The stemming porter is the process of converting the non-root words to root words. The roots words are the words which carries no emoticon emphasis. The emoticon emphasis is added to the word as emoticon weight age in the sentence according the grammar rules. The grammar rules sometimes increase or decrease the emphasis of the words being used in the sentences. In the existing work, the stemming porter has been used for the emotion mining, particularly sentiment analysis, on the data collected from social networks. The existing work is not very accurate in the terms of recall. The Recall value has gone lower while the authors were focusing on to improve the precision. The system effectiveness becomes higher for the sentiment analysis models, when the stemming porter is used and produces the results with higher accuracy measured in terms of precision and recall.

XIII.CONCLUSION

Here after the stemming process done in sentimental words in tweets in social media. The Sentiment analysis categories the positive words trimmed and also negative and neutral words also trimmed to find normal words in tweets, in future the smiley's word explanation also taken to the process of stemming

REFERENCES

[1]. Md. Abdur Rahman & Heung-Nam Kim &Abdulmotaleb El Saddik &Wail Gueaieb, A context-aware multimedia framework toward personal social network services, International Journal in Multimedia Tools Applications (2014) 71:1717–1747

[2]. Ahmed Al-Dhanhani, Rabeb Mizouni, Hadi Otrok, Ahmad Al- Rubaie, Analysis of collaborative learning in social network sites used in education, International journal of Social Network Analysis and Mining. (2015) 5:65

[3]. Anuradha Goswami, Ajey Kumar, Challenges in the Analysis of Online Social Networks: A Data Collection Tool Perspective, Wireless Personal Communications 18 August 2017,97:4015–4061

[4]. Cheng-Chung Chen · Xiaoxi Fu · Che-Yuan Chang, A terms mining and clustering technique for surveying network and content analysis of academic groups exploration, International Journal of Cluster Computing (2017) 20:43–52

[5]. E. Fersini1 · F. A. Pozzi1 · E. Messina1, Approval network: a novel approach for sentiment analysis in social networks, International journal of World Wide Web (2017) 20:831–854

[6]. Goran Putnik, Eric Cost, Ca´tia Alves, He´lio Castro, Leonilde Varela, Vaibhav Shah, Analysing the correlation between social network analysis measures and performance of students in social network-based engineering education, International Journal of Technology and Design Education, (2016) 26:413–437

[7]. Jong-Soo Sohn · Un-Bong Bae · In-Jeong Chung, Contents Recommendation Method Using Social Network Analysis,international journal of Wireless Personal Communications (2013) 73:1529–1546

[8]. Kawaljeet Kaur Kapoor & Kuttimani Tamilmani & Nripendra P. Rana2 & Pushp Patil &Yogesh K. Dwivedi & Sridhar Nerur , Advances in Social Media Research: Past, Present and Future, Information Systems Frontiers - Springer Nov 6, 2017

[9]. S. Venkatesan, Vladimir A. Oleshchuk. C. Chellappan, Sourabh Prakash, Analysis of key management protocols for social networks, , International Journal of Social Network Analysis and Mining. (2016) 6:3

[10]. Vishal A. Kharde, S.S. Sonawane,, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016