Sentiment Analysis Text POS Tagging on Movie reviews using NLTK

Akshaya R. Garje[#1],K. V. Kale[#2]

[#] *Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad - 431004, Maharashtra, India*

[1]akshayagarje33@gmail.com,[2]kvkale91@gmail.com

***Abstract*— User opinions or reviews are nothing but user generated content, and these are in huge number on the web that represents current form of user's feedback. Sentiment analysis is nothing but classifying these reviews into either positive or negative. Part of Speech (POS) Tagging can be applied by various tools and in different programming languages. This paper highlights pre-processing of data corpus and a tagging method which models directly to tag sparsity with other properties of POS tag assignments. This work mainly focuses on the Natural Language Toolkit (NLTK) library in the Python environment. The application of generated results helps to make review with accurate classification.**

***Keywords*— Sentiment analysis, NLTK, Pre-processing, POS tagging.**

I. Introduction

Large amount of textual data is available on web for different means of purposes. It increases its volume and complexity by means of narration and content, which found very difficult to mine for a specific task. It also increases time complexity to do it manually. Therefore, there is apparent problem in automatic categorization and organizing data.

Textual data contains facts and opinions. Facts focus on objective data transmission whereas the opinion expresses the author's sentiments. Initially, it was focused on categorization of the factual data. But nowadays we have number of websites through which we can contribute, modify or grade the content. Users can express their personal opinions on particular topics through blogs, forums, product review sites, and social networks [1].

Sentiment analysis is also known as opinion mining. It analyses opinions, sentiments, evaluations, appraisals, attitudes, and emotions which are related to products, services, organizations, individuals, etc. It mainly focuses on opinions which express positive or negative sentiments [2].

A. Sentiment analysis

It is a part of Natural language processing, text analysis, computational linguistics, and biometrics. It leads to systematically identification, extraction, quantification and subjective information. It is widely applied to voice of reviews and opinions on online and social media, and materials for application that range from marketing to customer services to clinical medicines [3].

Opinion words are one main focused difficult part of sentiment analysis. These are treated as positive for one side and may be negative for other one. Another part is of way of expressing the situation may vary from person to person. As per traditional text processing consideration, meaning does not change if there is a small change in few words of text, but in sentiment analysis it matters. For example, 'the book is good" is differs from 'the book is not good". The system analyses one sentence at a time because most of sentences have both positive and negative opinions.in some cases, statements are easy to understand by human mind but not by the system. "The movie was as good as its last movie", in this example, it is dependent on the information of another entity which data is not available [4].

B. Analysis using NLTK

NLTK was created in 2001 as part of a computational linguistics course in the Department of Computer and Information Science at the University of Pennsylvania. Since then it has been developed and expanded with the help of contributors. It has now been adopted in courses in dozens of universities, and serves as the basis of many research projects.

NLTK is a leading platform that is used for building Python programs to work with human language data. It provides friendly interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum [5].

II. Literature Review

Zhao Jianqiang and Gui Xiaolin, they evaluated the effects of different preprocessing methods in sentiment classification that includes removing URLs, replacing negation, reverting repeated letters, removing stop words, removing numbers and expanding acronyms. In that they used two feature model and four classifiers to identify tweet sentiment polarity on five twitter datasets. It shows results that the performance of sentiment classification improves after replacing negation and barely changes after removing URLs and stop words or numbers [6].

Emma Haddia, XiaohuiLiua, Yong Shib, they explored the role of text preprocessing in sentiment analysis and their results gives appropriate feature selection and representation and the resulted accuracies using support vector machines (SVM) can be improved up to the level achieved in topic categorization, often considered to be an easier problem [7].

R. Akila,R. Praveena Priya Darsini, they proposed a method that is efficient for pre-processing of tweets, that is important for classification. They performed three steps in pre-processing that are to remove URLs, to remove special characters and to tokenize the sentences. Here, NLP tokenizer was used for converting twitter dataset into database. In tokenization, they split the tweets to meaningful sentences. After that they pre-processed document which were converted to dataset that can be given to any machine learning algorithms [8].

K. Horecki and J. Mazurkiewicz, their idea was to divide text filtering process into three main parts known as later text filtering. It includes each level with additional techniques that gives better results than others. It is essential step of categorization as it determines categorization accuracy due to interrelation between information and noise amount that may have great influence on the results. Three level are descripted in following: Level 1 is Filtering based on short words, stop words and punctuation marks removal and also conversion to lowercase. It also includes lemmatization techniques. Level 2 is Filtering based on semantic trees analysis as well as removal of similar words. Level 3 is Filtering based on removal of adjectives and replacing them with corresponding nouns [9].

Deepak Singh Tomar et al, sentiwordnet is used based on algorithm that facilitates to identify the opinion nature i.e. positive or negative. Here, POS tagger which are adjective, nouns, adverb, etc., played vital role in finding out the accurate polarity. These POS taggers have their own abbreviations, so according to its tagger each word of sentence are specified by their abbreviation. If the POS taggers are in even count of numbers then its treated as positive and if its count is odd then its treated as negative [10].

Lluís Màrquez,Lluís Padró,Horacio Rodríguez have applied the inductive learning of statistical decision trees and relaxation labelling to the Natural Language Processing (NLP) task of morph syntactic disambiguation (Part Of Speech Tagging).The learning process is supervised and which obtains a language model oriented to resolve POS ambiguities, consisting of a set of statistical decision trees expressing distribution of tags and words in some relevant contexts. The acquired decision trees is directly used in a tagger which is both relatively simple and fast, and that has been tested and evaluated on the Wall Street Journal (WSJ) corpus with competitive accuracy [11].

Yuan Tian and David Lo, compared the effectiveness of seven state of the art POS taggers on bug reports. It build a ground truth set that contains 21,713 tagged words from 100 sampled bug reports from Eclipse and Mozilla project. Its preliminary experiment results shows that the state of the art POS taggers could achieve a reasonable accuracy on bug reports (83.6%-90.5%), although worse than its accuracy on a regular English corpus (97%, for most taggers) [12].

Mohamed Outahajala, Yassine Benajiba, Paolo Rosso, and Lahbib, the paper presents the first Amazighe POS tagger. There are very few resources that have been developed for Amazighe. The essential step needed for automatic text processing is the development of a POS tagger tool. The dataset used here have been manually collected and the 10-fold technique is used to further validate our results [13].

The literature survey carried for this work showed that pre-processing of the text for sentiment analysis plays an important role. Researchers have performed sentiment analysis for tweets, product reviews, news reviews, movie reviews and other similar text based media. Researchers have used various method in the pre-processing for sentiment analysis like removal of URLs, stop word removal, removal of special characters, POS tagging, tokenization, text filtering using toolkits like NLTK, POS tagger and similar toolkits developed by the researchers for NLP, text mining and sentiment analysis.

The researchers have shown that the output of pre-processing can be passed as an input to machine learning algorithms which gives better results for sentiment analysis. The accuracy varies depending upon the type of dataset used, pre-processing applied and the machine learning algorithm applied.

Researchers around the world are trying to implement machine learning algorithms for sentiment analysis or opinion mining for increasing the accuracy on different datasets using different approaches which is helping to increase the knowledge in this new area of research i.e. sentiment analysis and opinion mining which can be implemented in different application areas.

III. Methodology

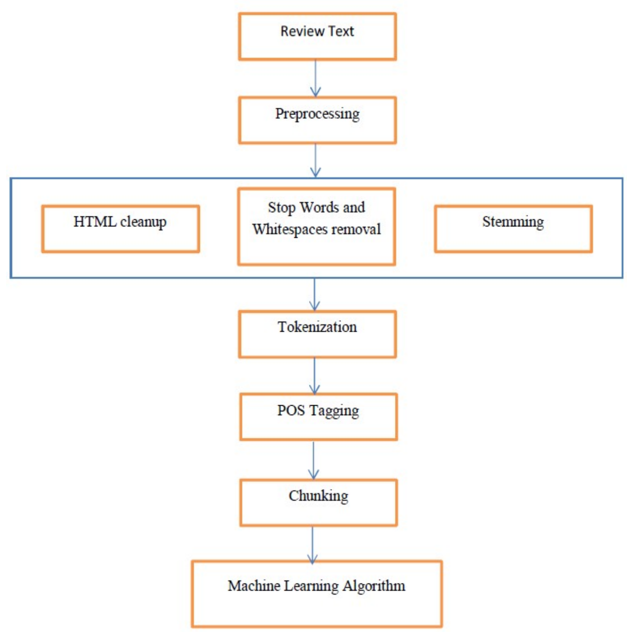

Fig. 1 Proposed methodology

A. *Preprocessing:*

Pre-processing of the data includes cleaning and preparing the data for classification process. It is necessary to preprocess the data because online texts usually has noise and uninformative parts namely, HTML tags, scripts, advertisements, etc. Many words do not impact the general orientation of the text. Each word is treated as the dimension of the text. The dimensionality increases by keeping irrelevant words in the text and hence more difficult is the classification process, these difficulties manifest in computational complexity of classification process as well as robustness of the analysis [14].

1) HTML cleanup:

Web pages contain irrelevant information such as HTML tags which are organized in different object elements so called <div> tags. To retain only the information of interest, this irrelevant information in the text should be cleaned. This cleaning also reduces the efficiency problems that arise due to this irrelevant information. In order to extract relevant information, there are many ways such as "HTML Cleanup", or the document object model (DOM), or the Apache library HTML Unit which parses the HTML specific features from texts. It also arranges them in an object-based tree structure that will be distinguished and separated from each other [11].

2) Stop words and whitespaces removal:

After HTML tags removal, some parts of text have two whitespaces. Hence, the process of removing one of those spaces for each occurrence of two spaces is known as Whitespace removal. And those words i.e. Stopwords which have no discriminant value in the text, or do not add any information to the general orientation of the text in terms of sentiment classification. These both things has to be removed, because due to their existence leads to less accurate results and longer processing time as there is increase in text dimensionality without additional data [14].

3) Stemming:

If a text contains the words, write, written and writing, they are treated as same words especially in a sentiment classifier where they have the same meaning and the same polarity. Stemming is the technique used in preprocessing the data which deletes the suffix of the words and returns the basic form of word. By stemming this text it will return them to "write" and then the word's frequency will be 3, and that is instead of three words with a frequency 1. Stemming helps to reduce dimensionality as well as to correctly identify words weights and importance in a text through their frequencies [14].

B. Tokenization:

Tokenization is the process of breaking a series of text into meaningful words, phrases or symbols. These meaningful elements are known as tokens, which can be used further for parsing. Generally tokenization is considered to be easy relative to other tasks in text mining.

Each word is a token when a sentence is tokenized into words whereas each sentence is a token if any paragraph is tokenized into sentences. Here, wordpunct_tokenizer(), word_tokenize(), sent_tokenize() are the methods of NLTK that are used [15].

C. POS Tagging:

Each word of sentence has its syntactic role which defines how the word is used. These syntactic roles are part of speech related to that word. In English, there are eight parts of speech which are known as the verb, the noun, the *pronoun, the* adjective, the adverb, the preposition, the conjunction and the interjection.

TABLE I.
PART-OF-SPEECH TAGS FOR VERBS

Tag	Definition
VB	base form
VBP	present tense not 3rd person singular
VBZ	present tense 3rd person singular
VBD	past tense
VBG	present participle
VBN	past participle

In natural language processing, part-of-speech (POS) taggers are used to classify words based on this parts of speech related to them. POS taggers are used in sentiment analysis due to two reasons: First, words such as nouns and pronouns can be filter out by POS tagger as they do not have any sentiment. Second, it helps in distinguishing words that can be used as different parts of speech. The tagger provides 46 different tags that can identify more detailed syntactic roles than only 8.Here, default tagger of NLTK is used for tagging. Table 3.3.1 lists all tags for verbs in the POS tagger [16].

D. Chunking:

As POS tagging is done research proceed to chunking. Chunking is the process to group words into meaningful chunks. Main goal of chunking here is to group words that are adjectives, verbs or adverbs.

In order to chunk the words, part of speech tags are combined with regular expressions here. From regular expression, following expressions are used:

+ = match 1 or more.
? = match 0 or 1 repetitions.
* = match 0 or more repetitions.
. = any character except a new line [17].

IV. EXPERIMENTAL WORK

In this work, movie review dataset is used from the nltk corpora, which contains 1000 positive and 1000 negative reviews. The snapshot of the dataset used is shown in figure 2. This dataset is been preprocessed which includes stop words and whitespaces removal, stemming, etc. which gives output as shown in figure 3.

Fig. 2. Overview of reviews in dataset.

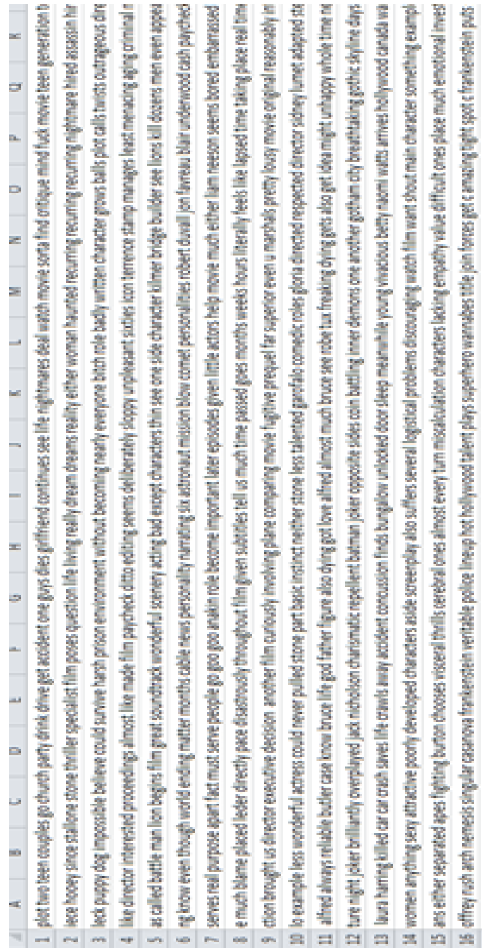

Fig. 3 Preprocessed text from dataset

After preprocessing and stemming the data, it was tokenized and then by using NLTK, it was POS tagged and then the processed data used to looks as shown in Table II.

Table II.
POS tagged words in dataset

WORDS	POS Tag
Plot	NN
two	CD
teen	NN
couples	NNS
go	VBP
church	NN
party	NN
drink	VBP
drive	JJ
get	NN
accident	JJ
one	CD
guys	NN
dies	VBZ

V. CONCLUSION

NLTK is the Python library which is mainly used in this research. The dataset is collected from the NLTK corpora itself. Various filters from NLTK library such as Stop words remover, tokenizer methods are used for preprocessing which results to clean text processing.

The Default POS tagger in the NLTK Library is applied to the dataset which is a simple and effective POS tagger of NLTK. This results various parts of speech features to the tokenized data. This POS tagger is comparatively better as it includes various types of Nouns, Pronouns, Adjectives and Adverbs according to their tenses which help the feature extraction of text processing to give more accurate results. The results accuracy for this dataset ranges from 81% to 85%. So this newly converted dataset is ready to be given as input to any machine learning algorithms for further sentiment classification.

ACKNOWLEDGMENT

The Authors acknowledge and extend our gratitude to UGC who have supported for development of UGC SAP (II) DRS Phase-I F.No.-3-42/2009 to Department of Computer Science & IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. Thanks to research supervisor for technical discussion and support at every stage of implementation.

REFERENCES

[1] Yessenov, Kuat, and Saša Misailovic. "Sentiment analysis of movie review comments." Methodology 17 (2009): 1-7.

[2] Yadav, Jyotika. "A Survey on Sentiment Classification of Movie." (2014).

[3] Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and Trends® in Information Retrieval 2.1–2 (2008): 1-135.

[4] Kalaivani, P., and K. L. Shunmuganathan. "Sentiment classification of movie reviews by supervised machine learning approaches." Indian Journal of Computer Science and Engineering 4.4 (2013): 285-292.

[5] Wagner, W. (2010). Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit. Language Resources and Evaluation, 44(4), 421-424 (accessed on 28 june 2017).

[6] Jianqiang, Zhao, and Gui Xiaolin. "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis." IEEE Access 5 (2017):2870-2879.

[7] Haddi, Emma, Xiaohui Liu, and Yong Shi. "The role of text pre-processing in sentiment analysis." Procedia Computer Science 17 (2013): 26-32.

[8] R. Akila,R. Praveena Priya Darsini, 'Twitter Data Preprocessing Using Natural Language Processing', South Asian Journal of Engineering and Technology, Vol.3, No.9 (2017) 46–49

[9] Horecki, Krystian, and Jacek Mazurkiewicz. "Natural Language Processing Methods Used for Automatic Prediction Mechanism of Related Phenomenon." International Conference on Artificial Intelligence and Soft Computing. Springer, Cham, 2015.

[10] Deepak Singh Tomar et al, 'A Text Polarity Analysis Using Sentiwordnet Based an Algorithm' (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 190-193

[11] Màrquez, Lluís, Lluis Padro, and Horacio Rodriguez. "A machine learning approach to POS tagging." Machine Learning 39.1 (2000): 59-91.

[12] Tian, Yuan, and David Lo. "A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports." Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on. IEEE, 2015.

[13] Outahajala, Mohamed, et al. "Pos tagging in Amazighe using support vector machines and conditional random fields." International Conference on Application of Natural Language to Information Systems. Springer, Berlin, Heidelberg, 2011.

[14] Haddi, Emma. Sentiment analysis: text, pre-processing, reader views and cross domains. Diss. Brunel University London, 2015.

[15] Vishwanathan, Shravan. "Sentiment Analysis for Movie Reviews." Proceedings of 3rd IRF International Conference, 10th May-2014, Goa, India. 2010.

[16] Fang, Xing, and Justin Zhan. "Sentiment analysis using product review data." Journal of Big Data 2.1 (2015): 5.

[17] Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006.

[18] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." LREc. Vol. 10. No. 2010. 2010.

[19] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." Proceedings of the workshop on languages in social media. Association for Computational Linguistics, 2011.

[20] Feldman, Ronen. "Techniques and applications for sentiment analysis." Communications of the ACM 56.4 (2013): 82-89.

[21] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11.538-541 (2011): 164.