"Sentiment Analysis : Datamation in Python and Weka"

Ms. Puja M. Dadhe[1], Dr. R.N. Jugele[2] and Mr. D.S. Sadhankar[3]
[1] Research Scholar, Department of Computer Science,
Shivaji Science College, Nagpur.
[2]Assosiate Professor, Department of Computer Science,
Shivaji Science College, Nagpur.
[3]Assistant Professor, Department of Computer Science,
SFS College, Nagpur.
poojadadhe@gmail.com[1], rn_jugele@yahoo.com[2],dileep.sadhankar@gmail.com[3]

Abstract: *Data analysis is a process of investigating and analysing data in order to derive some useful information. With the increase in use of internet lot of data is generated every day. About 80 percent of this data is unstructured, which needs a proper measure to be analysed. The objective of such analysis has a wide scope in discovering interesting information in the areas like business, politics, research, science and social science domains. Sentiment analysis plays a very important role in decision making process. It classifies a document, sentences and aspect in to positive and negative sentiments. It employs Supervised and unsupervised approaches to find polarity. The paper presents a comparison between Weka and Python as a tool for sentiment analysis on different datasets and compares the accuracy for each dataset.*

Keywords: **Sentiment analysis, Datasets, Weka, Python, Movie Reviews, Accuracy, Sentiment polarity, NLP, Naïve Bayes, Negative, Positive.**

1. INTRODUCTION

Sentiment Analysis is a branch of Natural language processing which deals with the problem of classification and a method of computing and satisfying a view of a person given in a piece of a text, to identify persons thinking about any topic is positive negative or neutral [4]. It is done on the data that is collected from the Internet and various social media platforms. Organizations, Companies and Governments often use sentiment analysis to understand how the people feel about themselves, products and their policies. The purpose of Sentiment analysis is to classify the polarity of user's sentiment and extract his opinion regarding an interested entity, which help in providing valuable information for decision making [3]. Polarity in sentiment analysis means classifying the sentiments as positive, negative and neutral. Further it is classified into different levels:

- **Document level**: This classifies the whole document text into positive or negative polarity.
- **Sentence level**: This extracts the polarity of each sentence of a document into positive or negative polarity.
- **Aspect/entity level**: This classifies the sentiment polarity of each entity's aspect or feature of a sentence/document [3].

With social media becoming the main platform for expressing feelings, views of a person now days, it is also gaining a lot of exposure for finding credible information. Social media refers to websites and application that are designed to allow people to share content quickly, efficiently and in real time[6]. Twitter and Facebook are most common social media

platforms. In a scenario where analytics uses various methods for data analysis, The primary goal of data analytics is to help companies make more informed business decisions. It is performed by enabling data scientists, predictive modellers and other analytics professionals to analyze large volumes of transaction data. The other forms of data is untapped by conventional Business Intelligence (BI) programs which include Web server logs, Internet clickstream data, social media content, social network activity reports, text from customer emails, survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet [6]. Social data plays an important role in online trading, as it depends on the stakeholder‘s interest. The comments or the opinion from the stakeholders play multiple roles. Sentiment analysis or opinion mining has been exploited to process this information. The process of discovering the subjective information using natural language processing, text analysis, and computational linguistics from the social data is known as sentiment analysis [1]. The Data sets used in this paper consists of views of customers, viewers and data collected from social media.

There are two types of techniques used in Sentiment Analysis:

- Machine learning based techniques: Here various machine learning algorithms like Naïve Bayes, Maximum Entropy, SVM, K-means etc are used for classification of sentiments. It plays an important role in designing a tool. Various supervised and unsupervised machine learning algorithms can be used for finding the sentiment analysis [5].
- Lexicon Based technique: In Lexicon, a sentiment dictionary is used with sentiment words for classification of sentiments.

In this paper, the technique Naïve Bayes a has been evaluated for finding accuracy, precision and recall. Naïve Bayes is taken because Naive Bayes is a high bias, low variance classifier, also it can build good model with small data set.

Lots of free and open source tools are available online like NLTK, Weka, Python, Rapid miner, GATE, Open NLP etc. In this paper two tools, Weka and Python has been used to analyse the sentiments collected from different datasets. Weka is a open source software which encompasses data analysis, data integration and reporting in a single suit. It is very easy to use software with lots of features like cross validation, performance vector, split validation. Weka is easy to use with friendly interface. The reason for choosing this as a tool for Sentiment Analysis is due to its GUI and ready to use properties. Python is general purpose and high level programming language use for developing desktop GUI applications. This paper analyses Accuracy of four Datasets employing Naïve Bayes in Weka and Python.

2. DATA SOURCE AND DATA SETS

Following are the data sets used for performing Sentiment Analysis in Weka and Python.

- Dataset1- The first data set used is movie data known as sentiment polarity dataset downloaded from http://www.cs.cornell.edu/people/pabo/movie-review-data. This

dataset contains two data files, pos and neg. each file contains 5000 positive and negative statements respectively. Sentence level sentiment analysis is done on it.

- Dataset2- Second dataset is Next Data set is also movie review data set which contains 1000 positive and 1000 negative text files downloaded from www.Kaggle.com
- Dataset3- Third data set used is imdb-sentiment-2011. This is large dataset consisting of 25000 positive and 25000 negative movie reviews. Downloaded from ai.standford.edu. This dataset is in .arff form. For python it is converted into .csv format.
- Dataset4- Fourth data set used is tf.data.dataset by tensoreflow. It consists of 900 postive and 900 negative files each.

3. METHADOLOGY

The main aim of this research is to analyze the accuracy of sentiments polarity using Naive Bayes technique also comparison between these techniques in Weka and Python to find out the best performing one. Diagrammatical representations of process involve in sentiment analysis in both Weka and Python is given below.

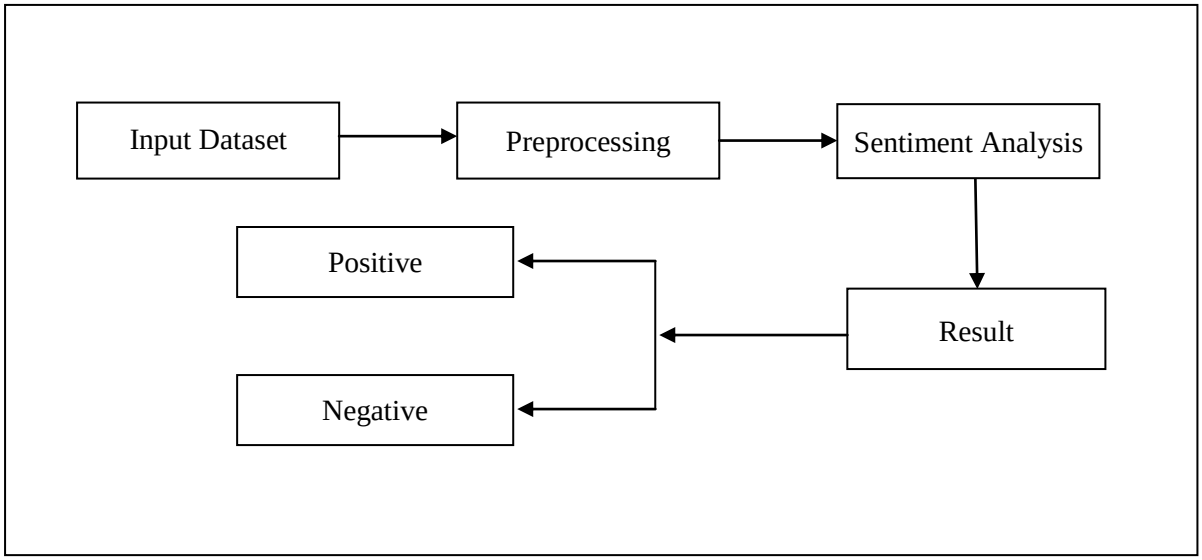

Fig 1: Sentiment Analysis Process

A preliminary Pre-Processing phase and attribute selection is essential for the sentiment classification task to be done which involves the following

- **Text preprocessing and feature extraction:** For the classification task to be done, a preliminary phase of text preprocessing and feature extraction is essential. To build the vocabulary, various operations are typically performed [2].

- **Word parsing and tokenization**: In this phase, each document is analyzed with the purpose of extracting the terms. Separator characters must be defined, along with a

tokenization strategy for particular cases such as accented words, hyphenated words, acronyms, etc [2].

- **Stop-words removal**: A very common technique is the elimination of frequent usage words: conjunctions, prepositions, base verbs, etc [2].
- **Lemmatization:** The lemmatization of a word is the process of determining its lemma. The lemma can be thought of as the "common root" of various related inflectional forms for instance, the words walk, walking and walked all derive from the lemma walk [2].
- **Stemming:** A simple technique for approximated lemmatization is the stemming. It works by removing the suffix of the word, according to some grammatical rules [2].
- **Term selection/feature extraction:** The term set resulting from the previous phases has still to be filtered, since we need to remove the terms that have poor prediction ability (w.r.t the document class) or are strongly correlated to other terms [2].

Paper focus on the **Naïve Bayes**, machine learning technique for both the tools.

Naïve Bayes: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. This model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. It is a simple probabilistic classifier based on Bayes' theorem which can build a good model even with a small data set. It is simple to use, computationally inexpensive and is very useful for the case where dimensions of input are high and for a given class as positive or negative, the words are conditionally independent of each other [7].

Naïve Bayes classifier [5] is an approach in which a classification of text (specific attribute) on the bases of appearance or absence of a class c in a given document d.

$$P(c|d) = \frac{\mathrm{P(c)P(d|c)}}{\mathrm{P(d)}}$$

Where c belongs to the positive or negative class and d belongs to the document whose class is being predicted, also P(c) and P(d|c) obtained during training.

To calculate accuracy of datasets following are some key terminology used which includes

Confusion matrix – It is also known as error matrix, is required to compute the ***accuracy*** of the machine learning algorithm in classifying the data into its corresponding labels. Confusion matrix C is a square matrix where Cij represents the number of data instances

which are known to be in group i(true label) and predicted to be in group j(predicted label) [8].

If we consider a binary classification problem,

Coo represents the count of true negative

Coi represents the count of false positive

Cio represents the count of false negative

Cu represents the count of true positive.

Represented as

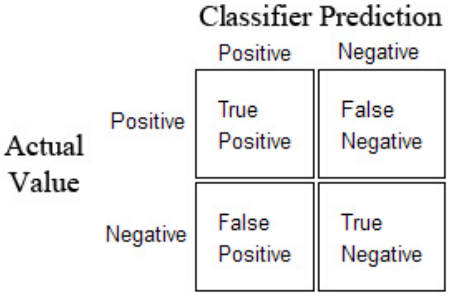

Fig 2: Confusion Matrix

Accuracy represents the number of correctly classified data instances over the total number of data instances calculated as-

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN}$$

To get perfect Accuracy following parameters are considered-

Precision refered as positive predictive value calculated as

$$\text{Precision} = \frac{TP}{TP+FP}$$

Precision should ideally be 1 (high) for a good classifier. Precision becomes 1 only when the numerator and denominator are equal i.e $TP = TP + FP$, this also means FP is zero. As FP increases the value of denominator becomes greater than the numerator and *precision* value decreases [8].

Recall is also known as sensitivity or true positive rate and is defined as follows:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Recall should ideally be 1 (high) for a good classifier. Recall becomes 1 only when the numerator and denominator are equal i.e $TP = TP + FN$, this also means FN is zero. As FN increases the value of denominator becomes greater than the numerator and recall value decreases [8].

So ideally in a good classifier, both precision and recall to be one which also means FP and FN are zero. Therefore we need a metric that takes into account both precision and recall. F1-score is a metric which takes into account both precision and recall and is defined as follows:

$$\text{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 Score becomes 1 only when precision and recall are both 1. F1 score becomes high only when both precision and recall are high. F1 score is the harmonic mean of precision and recall and is a better measure to find more perfect accuracy [8]. All these are evaluated on different dataset in weka and python and accuracy is been compared.

4. RESULTS

In this paper all four datasets were evaluated for accuracy in both the tools Weka and Python and following table shows the results procured by the tools.

Sr.no	Datasets	Positive/Negative files	Accuracy in weka	Accuracy in Python
1	Rt-polarity	5000	100%	77%
2	Txt-sentoken	1000	81.45%	78%
3	Imdb-sentiment-2011	25000	81.2%	82.%
4	Tf-data.dataset	900	81.33%	90.5%

Table 1: Results Showing the Accuracy values

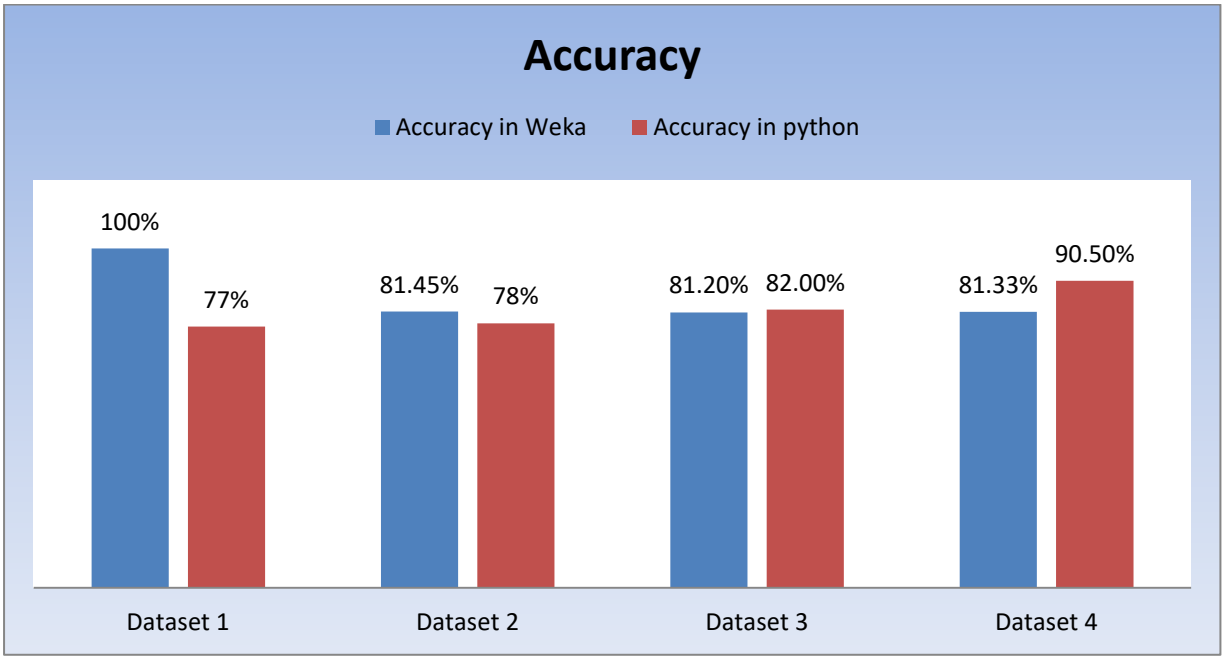

Chart 1: Comparison of Accuracy of Datasets in Weka and Python

5. CONCLUSION

With the Data set1 of 5000 text, Naïve Bayes in weka showed 100% accuracy whereas in Python it come out to be 77%, less compared to weka. For Dataset2 which consist of 1000 files again Weka Performed better than Python. But for Dataset3 and Dataset4 Python Accuracy is higher as compared to Weka. In Python processing is faster as compared to weka. Weka is little incompatible with large dataset. Time taken by Python to produce result is higher than Weka as observed while performing classification with both the tools. In this paper Naïve Bayes has been compared and observed for both the tools which show both the tools perform well for Sentiment Analysis.

In future these different data sets and methods can be taken to find out accuracy or comparison between different tools available. One can make use of rapid miner tool and R programming to find out how they work for sentiment Analysis. Also in this paper Naïve bayes has been used but the experiment can be also done with methods like SVM, Maximum Entropy etc. Instead of Datasets, Twitter data can be taken for classifying Sentiments.

6. REFRENCES

1. G. Ramajayam, 2Dr. V. Radhika, A Survey on Role Of Sentiment Analysis In Stakeholer's Statisfaction, International Journal of Pure and Applied Mathematics, Volume 119 No. 18 2018, 3021-3027.
2. Jincymol joseph, J R jeba, Information Extraction Using Tokenisation And Clustering Methods, International Journal of Recent Technology and Engineering(IJRTE), Volume-8 Issue-4,November 2019.
3. Ms. Puja M. Dadhe, Dr. R.N. Jugele, Inspection of Retrospection : Challenges of Sentiment Analysis, Compliance Engineering Journal,Volume 11,Issue 1,2020.
4. Rupinder Kaur et al, A Review on Sentimental Analysis on Facebook Comments by using Data Mining Technique International Journal of Computer Science and Mobile Computing, Vol.8 Issue.8, August- 2019, pg. 17-21.
5. Shilpa Singh Hanswal, Astha Pareek, Twitter Sentiment Analysis using Rapid Miner Tool, International Journal of Computer Applications (0975 – 8887) Volume 177 – No. 16, November 2019.
6. https://www.digitalvidya.com/blog/big-data-applications/
7. https://www.analyticsvidya.com/blog/2017/09/naive-bayes-explained/
8. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/