Sentiment Analysis using Machine Learning Techniques on Python

Nisha Rathee
Assistant Professor, Department of Information Technology
Indira Gandhi Delhi Technical University For Women
Delhi, India
nisharathee@igdtuw.ac.in

Nikita Joshi
Department of Information Technology
Indira Gandhi Delhi Technical University For Women
Delhi, India
joshi.nikita1396@gmail.com

Jaspreet Kaur
Department of Information Technology
Indira Gandhi Delhi Technical University For Women
Delhi, India
jaspreet.k.c1996@gmail.com

***Abstract*— Fundamentally, a sentiment refers to the reflection of emotions of people. Today's world stands on the strings of emotions. People express happiness, sadness, love, hatred etc. through some actions. Division of emotions i.e positive, neutral and negative, is called sentiment analysis. Nowadays there is a sentiment rich data in the form of tweets, status updates, blog posts, reviews, comments, forums for discussion etc. If we efficiently work upon this bucket full of sentiment rich data, it gives way in apprehending the opinions, views or perspective of masses in a specific functional area. Moreover, the result of this analysis will aid people in taking suitable actions or corrective measures for their growth. This effort of ours is like a drop in the ocean to try to analyze the reviews posted by people at four different websites (airlinequality.com, Amazon, Yelp, and IMDB). Further, the reviews are processed and analyzed using machine learning procedures, algorithms and other related aspects. Finally, the conclusion is derived by finding the polarity of a particular review whether it is poor, average or excellent for Airlines dataset and 0 or 1 for the other three datasets. The entire task was performed using Python.**

keywords: Sentiment Analysis, Python, Machine Learning, IMDB Reviews, Yelp, Indian Airlines, Amazon Reviews

I. INTRODUCTION

The era of internet has spread to such a level that 51% of the world's population had internet access as recorded on June 2017. The International Telecommunication Union stated that "out of the 51 % about 2 billion people are from developing countries and 89 million from least developed countries." [Fig 1] This audience of internet users has created an abundance of data which, if effectively processed and analyzed can prove to be really helpful.

Nowadays, reading and writing reviews play a major role in our whole shopping or buying process. More people are looking for reviews than ever before. From buying a novel to buying mobile phones, cars or even booking a hotel, one always checks for reviews and ratings that are submitted by other customers. According to Pew Research, "24% of the buyers have posted comments or reviews online about the products they buy." Research also demonstrates that "91% of the people periodically and , continually go through online reviews, and 84% trust online reviews as much as a personal recommendation" [2]. This review process plays a substantial role in molding our decision towards buying a particular product. [1] Yelp & Facebook followed by Google & BBB.org are the most trusted review sites for local consumers. From the above stats, it reveals that the consumers have turned to be more review-savvy, which is evident from their shopping patterns and behaviors.

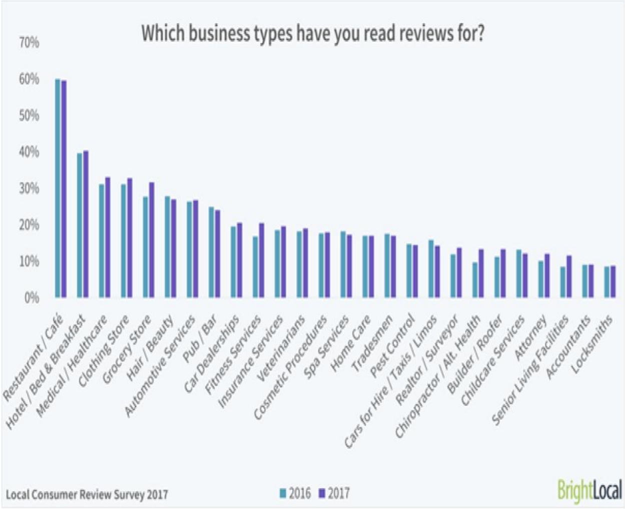


Fig. 1. Statistics of different types of reviews [1]

The principal task of Sentiment Analysis is to find the perspective, view, attitude or feelings of a speaker on a particular topic, event or interaction. Basically, it is the analysis of an emotionally charged text. [3] Fundamentally, it addresses the question, "what do people feel about a certain topic" and hence finding its polarity i.e dividing it into three classes: neutral, negative and positive, or a range of polarity like star ratings for a movie etc. Other associated terms used for Sentiment Analysis are Appraisal Attraction, Opinion Mining and Subjectivity Analysis.

Another very important aspect to be considered while dealing with sentiments is its two categories, implicit and explicit. The former one has a direct meaning (e.g. Yesterday was a beautiful day) and the latter one has a hidden meaning (e.g. the new machine broke down in a day). However, the analysis of the latter is a cumbersome process, due to its

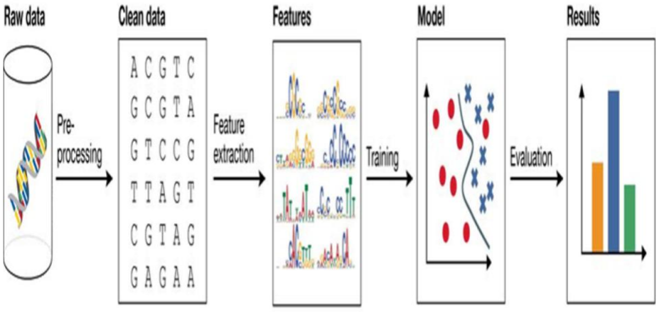
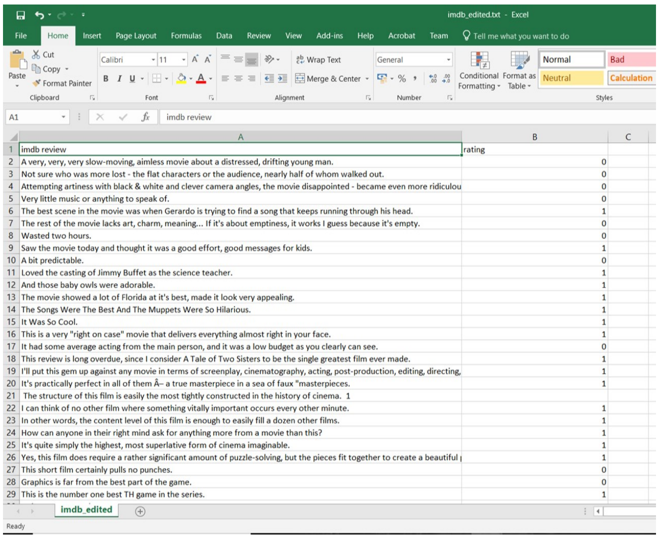


Fig. 2. Stages of Machine Learning[4]



	A	B	C
1	imdb review	rating	
2	A very, very, very slow-moving, aimless movie about a distressed, drifting young man.	0	
3	Not sure who was more lost - the flat characters or the audience, nearly half of whom walked out.	0	
4	Attempting artiness with black & white and clever camera angles, the movie disappointed - became even more ridiculou	0	
5	Very little music or anything to speak of.	0	
6	The best scene in the movie was when Gerardo is trying to find a song that keeps running through his head.	1	
7	The rest of the movie lacks art, charm, meaning... If it's about emptiness, it works I guess because it's empty.	0	
8	Wasted two hours.	0	
9	Saw the movie today and thought it was a good effort, good messages for kids.	1	
10	A bit predictable.	0	
11	Loved the casting of Jimmy Buffet as the science teacher.	1	
12	And those baby owls were adorable.	1	
13	The movie showed a lot of Florida at it's best, made it look very appealing.	1	
14	The Songs Were The Best And The Muppets Were So Hilarious.	1	
15	It Was So Cool.	1	
16	This is a very "right on case" movie that delivers everything almost right in your face.	1	
17	It had some average acting from the main person, and it was a low budget as you clearly can see.	0	
18	This review is long overdue, since I consider A Tale of Two Sisters to be the single greatest film ever made.	1	
19	I'll put this gem up against any movie in terms of screenplay, cinematography, acting, post-production, editing, directing,	1	
20	It's practically perfect in all of them Â– a true masterpiece in a sea of faux "masterpieces.	1	
21	The structure of this film is easily the most tightly constructed in the history of cinema. 1		
22	I can think of no other film where something vitally important occurs every other minute.	1	
23	In other words, the content level of this film is enough to easily fill a dozen other films.	1	
24	How can anyone in their right mind ask for anything more from a movie than this?	1	
25	It's quite simply the highest, most superlative form of cinema imaginable.	1	
26	Yes, this film does require a rather significant amount of puzzle-solving, but the pieces fit together to create a beautiful	1	
27	This short film certainly pulls no punches.	0	
28	Graphics is far from the best part of the game.	0	
29	This is the number one best TH game in the series.	1	

imdb_edited

Fig. 3. Screenshot of IMDB Dataset

not-so-direct meaning.

Another important concept used in sentiment analysis is data mining. Data mining is a technique of recognizing distinct, beneficial and intriguing patterns in huge amounts of data. Its varied applications are as follows:

- In Business Sector(banking, retail, insurance)
- In government security (search for terrorists and criminals) and
- In scientific research (medicine, astronomy)

It will also allow us to predict future possibilities of an enterprise, assist decision-making processes and aid formation of appropriate strategies.

The process of analysis involves identification of if/then combinations, followed by the application of support and confidence criteria. Support refers to the frequency of a particular item that occurs in a dataset, while the number of occurrences of if/then combinations, which are accurate determines the confidence.

Some other data mining parameters include Forecasting, Clustering, Path or Sequence Analysis and Classification.

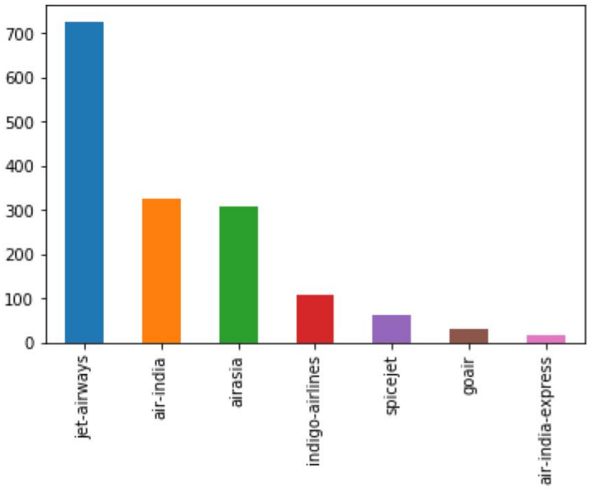


Fig. 4. Indian Airlines dataset distribution

- Forecasting parameters
  Additionally, through data mining, interesting patterns are discovered in our dataset that can help us to reasonably predict the future, also known as predictive analysis.
- Clustering parameters
  Placing a set of objects in a single bracket and their aggregation depending upon how closely related they are to each other.
- Path or Sequence Analysis
  Finds patterns in which one event is a consequence of other.
- Classification
  This parameter finds new patterns, that may transform the organization of data. It forecasts the value of variables which rely on other aspects within a dataset.

The implementation makes use of classification parameters and techniques. This involves the prediction of class labels for each of the instances in the test set. Classification task involves application of eight machine learning algorithms, namely, Logistic Regression, k-nearest neighbors, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier, Gaussian Naive Bayes and Bagging Classifiers.

Bagging Classifier has been used twice, once on taking Random Forest Classifier as the base and once considering Ada Boost Classifier.

II. LITERATURE REVIEW

In a very basic or elementary format, the meaning of sentiment analysis is well explained by Yelena Mejova[3].

Along with the objective of Sentiment Analysis, methodologies to tackle the text, mainly - Part-Of-Speech and Machine Learning, are also discussed considering their significance. In Sentiment Analysis, frequency vs term presence discussion points to the importance of finding the most unique words rather than the most frequent ones. Also,

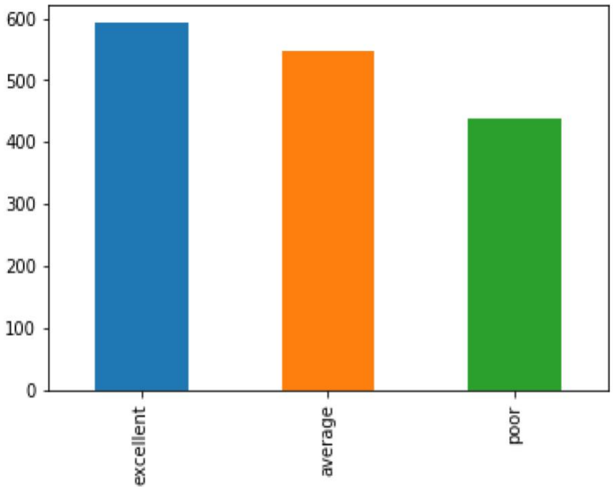


Fig. 5. Indian Airlines polarity distribution

negation plays a very vital role while accessing sentiments as it can invert the polarity of the whole phrase. We present a new feature vector for classifying the tweets as positive, negative and extract people's opinion about products.

Neethu M S et al. [5], performed sentiment analysis of tweets on electronic commodities such as mobiles, laptops etc by applying Machine Learning Algorithms. A new and distinct feature vector was proposed to categorize the tweets on the basis of their polarity as negative, positive and extraction of people's opinion about electronics products. In order to identify sentiments from tweet two techniques that can be implemented are Symbolic technique & Machine learning techniques, but the latter one was much better and efficient. Many classifiers like SVM, Maximum Entropy, Naves Bayes, and Ensemble are used to test the accuracy of the feature vector. The results of all these classifiers show that all of them have almost similar accuracy for the new feature vector.

Some problems predominating in sentiment research includes feature based classifications, handling negation and sentiment classification. Vinodhini et al.[6] present survey about the methods and techniques involved in Sentiment Analysis. The various challenges in this field are also discussed. A fact that is stated here is that Sentiment Analysis Classifiers are dependent on the topics being categorized.It is quite clear that neither classifiers consistently surpasses each other's performances. To overcome individual drawbacks and benefit from each other, different set of features and classifiers are united in a proficient way. This finally improves the performance of algorithms in sentiment classification.

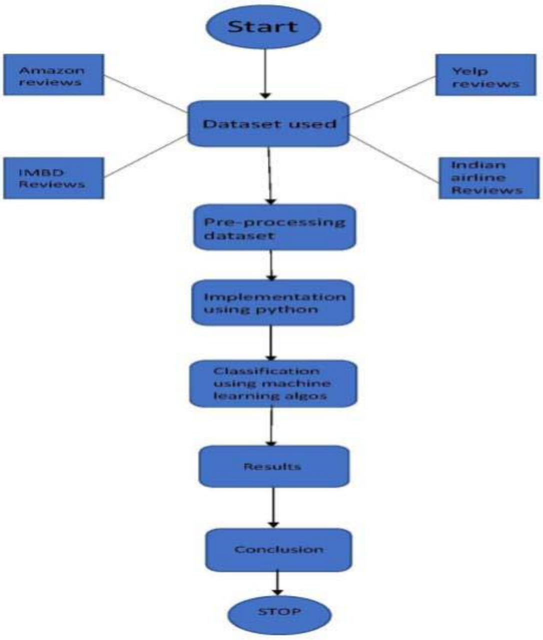


Fig. 6. Flowchart of events

Zhen Nui et al.[7] have presented an improved model which uses better ways for creating feature vector and calculation of weights. The paper points to the fact that how well known microblogging is in today's world and hence concentrates on its Sentiment Analysis. The steps which were followed in their analysis were: Extraction of keywords as feature items, Calculating the weights of each feature word then Training samples. After which Sentiment classification is done then the last step is evaluation of performance. Testing was done on Machine Learning Algorithms namely Nave Bayesian Text Classifier(NB), Maximum entropy (ME) and lastly on Support Vector Machine (SVM). Finally a new technique for sentiment analysis was introduced using Improved Bayesian Algorithm which can classify text with higher accuracy and efficiency and hence provide us with better results and analysis.

P. D. Turney[8] presented an unsupervised learning technique for categorizing a review as thumbs up (recommended) or down (not recommended). A review is considered as recommended if the average semantic orientation of the phrases is positive. The technique used has the following steps: (1) phrases having adjectives or adverbs are extracted, (2)semantic orientation of every phrase is estimated and (3) lastly classifying them depending upon the average semantic orientation of the phrases. While experimenting with 410 reviews dealing with banks, automobiles, travel destinations and movies from Epinions resulted that movie reviews are difficult to classify, than banks and automobiles reviews. Travel reviews are a middle level case.

III. PROPOSED METHODOLOGY

For our analysis, four different types of datasets are taken into consideration, namely: Amazon reviews, Yelp reviews, IMDB reviews [Fig 3] and Indian Airlines reviews [Fig 4] .

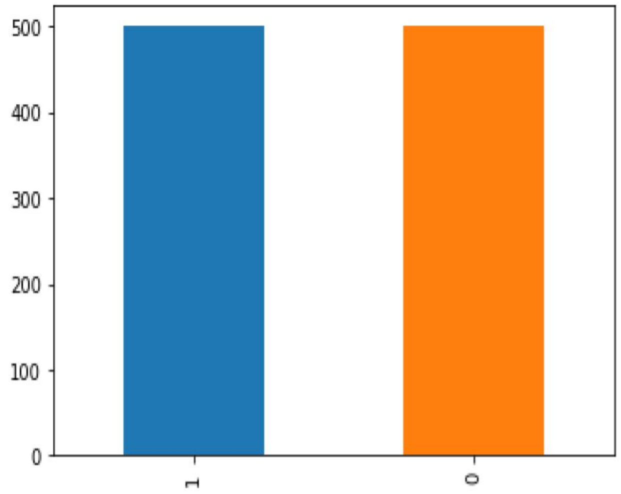


Fig. 7. IMDB,Amazon,Yelp polarity distribution

These datasets were pre-processed according to the prerequisites of the various algorithms used. The next step was to leverage Python and incorporate these datasets into machine learning models. Finally, varied accuracies were obtained for different algorithms in different datasets. The conclusive judgment is descended by specifying which algorithm works best or worst for which dataset. [Fig 14]

IV. IMPLEMENTATION

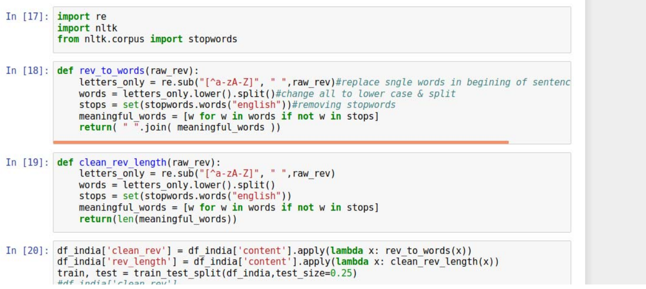
This section comprises of all the steps executed to accomplish the sentiment predictive goal in detail. The steps involve:- Formation of datasets, preprocessing of data, the creation of feature vector and finally the task of classification.

A. FORMATION OF DATASET

Our research is based on the of account four datasets for the task of sentiment analysis.
1.Indian-Airlines:- Actually, the dataset only for Indian airlines was not directly available, but a dataset is found that constitutes data of airlines emanating from the globe which can be traced from GitHub, https://github.com/quankiquanki/skytrax-reviews-dataset/tree/master/data Now, next task involved the extraction of reviews for Indian airlines. This was performed using a Python script. Presently, the updated dataset includes dataset of the seven Indian airlines: Jet-Airways, Air-India, Airasia, Indigo-airlines, SpiceJet, Goair, Air-India-Express.[Fig 4]

DISTRIBUTION OF CLASSES IN DIFFERENT DATASET

Two columns were employed for the implementation, ‘content’ (which contains the review) and ‘overall_rating’ (rating of an airline on the scale of 1 to 10). The overall_rating was further mapped to create an attribute ‘summary’ which comprises values of type: poor, average and excellent [Fig 5]. The other 3 datasets contain reviews from Amazon, Yelp and IMDB. These include two columns ‘review’ and ‘rating’. Rating attribute includes values of type 0 or 1, 0 for negative and 1 for positive [Fig 7]. This ‘review’ contains the actual content to be processed. All these four datasets have been taken from the UCI repository[9]. All these datasets have 1000 instances. All the four dataset files are read into the programs and are stored in data frames using ‘pandas’[11]. Pandas make it really easy to manage and manipulate data frames. ‘Matplotlib’ [12] is used for plotting all the graphs.


```
In [17]: import re
         import nltk
         from nltk.corpus import stopwords

In [18]: def rev_to_words(raw_rev):
             letters_only = re.sub("[^a-zA-Z]", " ",raw_rev)#replace sngle words in begining of sentenc
             words = letters_only.lower().split()#change all to lower case & split
             stops = set(stopwords.words("english"))#removing stopwords
             meaningful_words = [w for w in words if not w in stops]
             return( " ".join( meaningful_words ))

In [19]: def clean_rev_length(raw_rev):
             letters_only = re.sub("[^a-zA-Z]", " ",raw_rev)
             words = letters_only.lower().split()
             stops = set(stopwords.words("english"))
             meaningful_words = [w for w in words if not w in stops]
             return(len(meaningful_words))

In [20]: df_india['clean_rev'] = df_india['content'].apply(lambda x: rev_to_words(x))
         df_india['rev_length'] = df_india['content'].apply(lambda x: clean_rev_length(x))
         train, test = train_test_split(df_india,test_size=0.25)
         #df_india['clean_rev']
```

Fig. 8. Coding in Jupyter[10] notebook

B. DATA PREPROCESSING

Every component of a sentence is not needed for the Sentiment Analysis. This arises the need for the ‘review’ preprocessing. Our implementation uses inbuilt libraries, ‘re’(regular expression) and ‘nltk’ (Natural Language Tool Kit)[13]. This includes:-
a) Replacement of single characters at the beginning of sentences by spaces(like I, A etc.).
b) Converting all the text to lowercase.
c) Removal of Stopwords: Stopwords are the words which complete a sentence but reveal no or negligible emotion of the review/text. These include words like ‘must’, ‘ourselves’, ‘me’, ‘they’. Its list can be traced from https://gist.github.com/sebleier/554280. This phase ends by placing all the cleaned reviews in another data frame. This makes us ready to create feature vectors for each of the row of cleaned tweets. This step is inevitable as character data cannot be fed to a machine learning classifier.

C. CREATION OF FEATURE VECTOR

This step involves the conversion of reviews to a vector. The ‘feature_extraction.text.CountVectorizer()’ method of “sklearn” is used to serve the required purpose. This leads to the creation of separate column for every distinct word, occurring in all the reviews. The cell is marked 0 or 1, depending upon its occurrence in that particular review.

D. CLASSIFICATION

This phase involves, training our model for making predictions using eight machine learning algorithms for all the four datasets. This phase is broken down into the following steps -

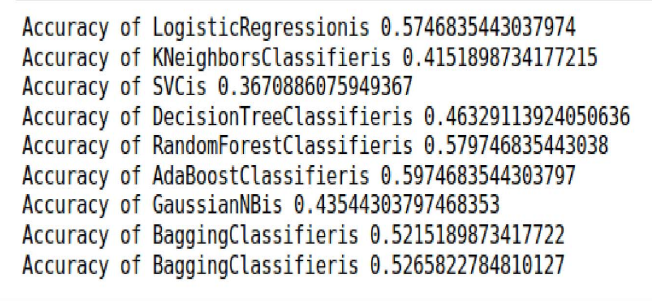
- **Train-Test Split** - The dataset is split into two parts: training set and testing set with split percentage as 75-25.

- **Cross-Validation** - It is one of the measures to prevent overfitting. Overfitting is a situation when the fitted model is overly complicated due to noise data-points in the training data. In such a case, the model gives a good accuracy on the training dataset but would perform worse on the unseen-data. In order to avoid such a situation, k-fold cross-validation is used to train the model.
  k-fold cross-validation partitions the data into k subsets. One of the sets out of k is reserved for the validation set and rest k-1 sets are used to train the model. Now, the left 1 set is what the model is tested against before making predictions for the test set. This process is repeated k times, and the accuracy is averaged out.[14]
  Firstly, the feature vectors along with their respective labels are fed as input to different classifiers. Then, the model trained on the input data is made to predict for another set of unseen values.
  Mathematically, if a feature vector 'X' and a label 'Y' are considered from a particular set, then the main task of classification would be to create a function C(X), that intakes the feature vector X and forecasts the value for Y.

- **Prediction** - This step involves the fitted model, from the prior step, to make prediction for the unseen set of data.
  All the models were trained using k-fold cross-validation. All the classifiers have been imported from the "sklearn" [15] module. [Fig 8]

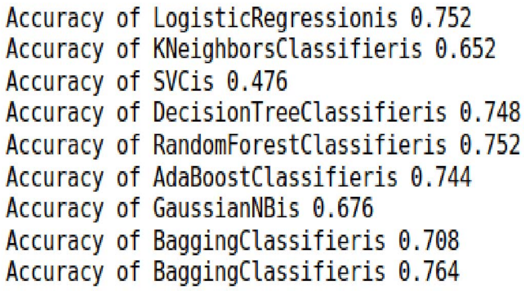
  1) **Logistic Regression** - The most preferred method for the classification of variables having two classes only. It is based on the equation comparable to linear regression. Based on the sigmoid function, input values (x) are united linearly using weights to forecast an output value (y). Equation for logistic regression,
  $$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$$
  where b_0 is the bias (intercept) and b_1 is the coefficient for a value x and y is the outcome/prediction. Each attribute/column in the data has an associated 'bi' value which is obtained by training the model. Then, this model is used for making predictions. This algorithm has been used by calling LogisticRegression() method of "sklearn".

  2) **k - Nearest Neighbors** - As the name suggests, "kNN" finds the nearest 'k' neighbors on the basis of similarity in their attributes. This algorithm considers the whole training dataset as its model. Whenever a prediction is requested, the model looks for its 'k' neighbors and then predicts the value of the unseen case on the basis of the summary of its neighbors. It is simple to use but is a "lazy learner". This algorithm has been used by calling KNeighborsClassifier() method of "sklearn".



```
Accuracy of LogisticRegressionis 0.5746835443037974
Accuracy of KNeighborsClassifieris 0.4151898734177215
Accuracy of SVCis 0.3670886075949367
Accuracy of DecisionTreeClassifieris 0.46329113924050636
Accuracy of RandomForestClassifieris 0.579746835443038
Accuracy of AdaBoostClassifieris 0.5974683544303797
Accuracy of GaussianNBis 0.43544303797468353
Accuracy of BaggingClassifieris 0.5215189873417722
Accuracy of BaggingClassifieris 0.5265822784810127
```

Fig. 9. Indian Airlines Results

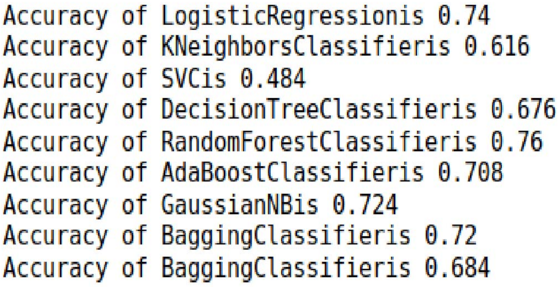


```
Accuracy of LogisticRegressionis 0.752
Accuracy of KNeighborsClassifieris 0.652
Accuracy of SVCis 0.476
Accuracy of DecisionTreeClassifieris 0.748
Accuracy of RandomForestClassifieris 0.752
Accuracy of AdaBoostClassifieris 0.744
Accuracy of GaussianNBis 0.676
Accuracy of BaggingClassifieris 0.708
Accuracy of BaggingClassifieris 0.764
```

Fig. 10. Amazon Results

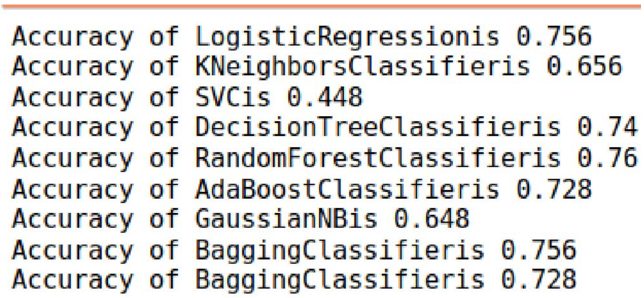
  3) **Support Vector Machine Classifier** - Support Vector Machine Classifier is applied to both classification and regression enigmas. They are known for their good performance. It uses a technique called the kernel trick, which typically computes the distance between two observations. This follows the search for decision boundary in order to find the distance between the closest members of separate classes. SVMs are robust in the cases of overfitting. This algorithm has been used by calling SVC() method of "sklearn".

  4) **Decision Tree Classifier** - Decision Tree Classifiers imitates the decision-making process as of humans. A tree is a collection of nodes, links (to the children nodes) and leaf nodes. Similarly, a decision tree is also a tree with each of its components with a little bit different interpretation. Every node represents an attribute. Reaching a child node involves going through a decision(link). Finally, leaf nodes represent the output. But, a decision tree is prone to overfitting data when trees have greater heights. A deep Decision Tree also has high-variance. This algorithm has been used by calling



```
Accuracy of LogisticRegressionis 0.74
Accuracy of KNeighborsClassifieris 0.616
Accuracy of SVCis 0.484
Accuracy of DecisionTreeClassifieris 0.676
Accuracy of RandomForestClassifieris 0.76
Accuracy of AdaBoostClassifieris 0.708
Accuracy of GaussianNBis 0.724
Accuracy of BaggingClassifieris 0.72
Accuracy of BaggingClassifieris 0.684
```

Fig. 11. IMDB Results



```
Accuracy of LogisticRegressionis 0.756
Accuracy of KNeighborsClassifieris 0.656
Accuracy of SVCis 0.448
Accuracy of DecisionTreeClassifieris 0.74
Accuracy of RandomForestClassifieris 0.76
Accuracy of AdaBoostClassifieris 0.728
Accuracy of GaussianNBis 0.648
Accuracy of BaggingClassifieris 0.756
Accuracy of BaggingClassifieris 0.728
```

Fig. 12. Yelp Results

DecisionTreeClassifier() method of “sklearn”.

5) **Random Forest Classifier** - This is an ensemble method which is based on the Decision Tree algorithm. Random forest creates a number of trees, unlike a decision tree algorithm which creates a single tree. The increase in the number of trees is directly proportional to an increase in the robustness of the algorithm. This algorithm overcomes the overfitting limitation of the Decision Tree algorithm along with reducing the biases. Random Forest is an algorithm which performs exceptionally well in most cases. This algorithm has been used by calling RandomForestClassifier() method of “sklearn”.

6) **AdaBoost Classifier** - AdaBoost Classifier is again an ensemble classifier, generally used for binary classification problems. Ada Boost is used to uplift the performance of weak classifiers. Decision trees with height 1 work really well with AdaBoost Classifier. Since these short trees involve only one decision to be taken and hence known as Decision stumps. These are prepared by allocating weight to all the rows of the training sample. Ada Boost being a binary classifier, requires only one decision to be taken by each stump. This algorithm has been used by calling AdaBoostClassifier() method of “sklearn”.

7) **Gaussian Naive Bayes** - Based on Naive Bayes, Gaussian Naive Bayes is used to handle real-time data with “continuous” distribution. It works on the assumption that input data follows Normal Distribution. The predictions are made by computing conditional probability of a particular class when its feature vector is given.

$$P(C_k|X(i)) = (P(X(i)|Ck) * P(C_k))/P(X(i))$$

where C_k-stands for a class and $X(i)$ represents a feature vector. The “naive” assumption is that, given a class, attributes are conditionally independent of each other. Gaussian NB is a highly efficient technique for classification. This algorithm has been used by calling GaussianNB() method of “sklearn”.

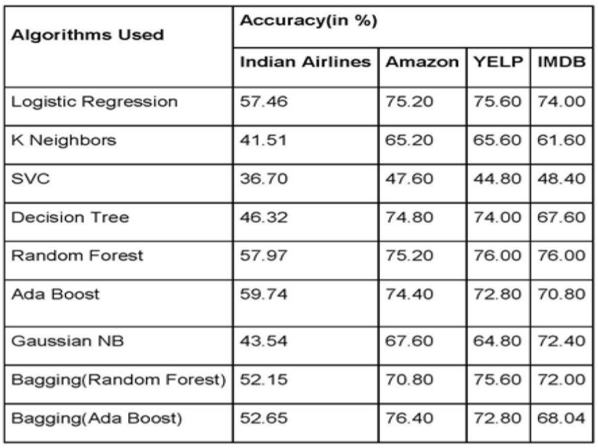
8) **Bagging Classifier** - A meta-heuristic, which is used to upgrade the performance of both regression algorithms and classification. It involves training on many models. Training set for each of the models is created by selecting a random sample of the whole training set. Testing is performed by taking the mean(regression) or voting(classification) of outcomes of all the algorithms. This technique is useful in case of algorithms that give higher variance or when there are a lesser number of instances in the dataset. The accuracy of Decision Tree is raised often by this method. This algorithm has been used by calling BaggingClassifier() method of “sklearn”.

V. RESULTS

After running all the above mentioned algorithms successfully on our datasets, the results are obtained as depicted by the table. [13] The variation in outcomes for Indian Airline reviews and the reviews from other three websites is due to the following reasons:
1. The difference in the length of reviews: Since reviews with longer lengths tend to add many new keywords to the columns when compared to the shorter reviews hence, a lesser probability of recurrence of those words in the unclassified reviews. Consequently, this factor gives rise to increased lengths of feature vectors and a poor accuracy.
2. The number of classes for classification: More classes for categorization decrease the probability of mapping a review into each class. And the reviews in the first dataset were classified into three classes while the reviews in the other datasets were categorized into two classes. This factor also contributed to the poor accuracy of results for Indian Airlines dataset.

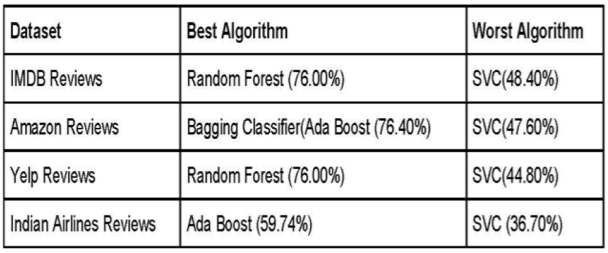
VI. CONCLUSION

The rich corpus of data set, helped us to find out various trends and patterns effectively. The variety of datasets, that were taken into our consideration assisted us to study and apply algorithms efficiently during the analysis. The analysis



Algorithms Used	Accuracy(in %)			
	Indian Airlines	Amazon	YELP	IMDB
Logistic Regression	57.46	75.20	75.60	74.00
K Neighbors	41.51	65.20	65.60	61.60
SVC	36.70	47.60	44.80	48.40
Decision Tree	46.32	74.80	74.00	67.60
Random Forest	57.97	75.20	76.00	76.00
Ada Boost	59.74	74.40	72.80	70.80
Gaussian NB	43.54	67.60	64.80	72.40
Bagging(Random Forest)	52.15	70.80	75.60	72.00
Bagging(Ada Boost)	52.65	76.40	72.80	68.04

Fig. 13. Results



Dataset	Best Algorithm	Worst Algorithm
IMDB Reviews	Random Forest (76.00%)	SVC(48.40%)
Amazon Reviews	Bagging Classifier(Ada Boost (76.40%)	SVC(47.60%)
Yelp Reviews	Random Forest (76.00%)	SVC(44.80%)
Indian Airlines Reviews	Ada Boost (59.74%)	SVC (36.70%)

Fig. 14. Conclusion

proved the important phenomenon correct that no algorithm can be ranked as best for sentiment analysis because these algorithms are domain specific. Some would work best for one type of data set and others would work best for some other type. Another important factor discovered was that the analysis of tweets is comparatively easier as compared to reviews due to the constraint of 140 character imposed by Twitter. The user has to express his/her entire emotion within these 140 characters whereas in reviews the user has one full comment section which can contain multiple paragraphs too. And it is quite obvious that analysis for sentiments in short texts would be simpler than paragraphs. This too is the reason for decreased accuracies of our algorithms. Finally, the best and worst algorithms in terms of accuracy for every dataset is specified in the table.[Fig 14] The field of sentiment analysis has a very wide scope of research and work. It helps to find the overall polarity of a huge amount of dataset in no time and the result can be used for further analysis, for growth, improvement, and betterment of that particular domain or sector.

REFERENCES

[1] Fig 1 Source: https://www.brightlocal.com/learn/local-consumer-review-survey/

[2] Source:https://www.inc.com/craig-bloem/84-percent-of-people-trust-online-reviews-as-much-.html

[3] Y. Mejova, “Sentiment analysis: An overview,” Comprehensive exam paper, available on http://www. cs. uiowa. edu/ ymejova/publications/CompsYelenaMejova. pdf [2010-02-03], 2009.

[4] Fig 2 Source:http://msb.embopress.org/content/12/7/878.

[5] Neethu M. S., Rajasree R Sentiment Analysis in Twitter using Machine Learning Techniques,4th ICCCNT 2013,July 2013, Tiruchengode, India.

[6] G. Vinodhini and R. Chandrasekaran, “Sentiment analysis and opinion mining: A survey,” International Journal, vol. 2, no. 6, 2012.

[7] Z. Niu, Z. Yin, and X. Kong, “Sentiment classification for microblog by machine learning,” in Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on, pp. 286289, IEEE, 2012.

[8] P. D. Turney, “Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, pp. 417424, Association for Computational Linguistics,2002.

[9] From Group to Individual Labels using Deep Features, Kotzias et. al,. KDD 2015.

[10] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Prez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damin Avila, Safia Abdalla, Carol Willing, Jupyter Development Team,cover Jupyter Notebooks a publishing format for reproducible computational workflows,87 - 90,,ElPub conference

[11] @InProceedings mckinney-proc-scipy-2010, Wes McKinney Data Structures for Statistical Computing in Python , Proceedings of the 9th Python in Science Conference 51 - 56 (2010), Stéfan van der Walt and Jarrod Millman.

[12] Hunter, J. D.Matplotlib: A 2D graphics environment, Computing In Science & Engineering,9,3,90–95,Matplotlib is a 2D graphics package used for Python for application development, interactive scripting, and publication-quality image generation across user interfaces and operating systems. IEEE COMPUTER SOC10.1109/MCSE.2007.55,2007

[13] Bird, Steven, Edward Loper and Ewan Klein , Natural Language Processing with Python. O’Reilly Media Inc.,2009.

[14] Fushiki T. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing 2011; 21 (02) 137-146.

[15] Scikit-learn: Machine Learning in Python, Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., JMLR 12, pp. 2825-2830,2011. 10.3233/978-1-61499-649-1-87,2016.