GRINREY

Advanced Research in Computer Engineering

Sandip A. Kale
Editor

Research Transcripts in Computer, Electrical and Electronics Engineering

5

Application of Classifiers for Assortment of Online Reviews

Biplob Kumar[a,*], Pritom Sarker[a] and Nakib Aman Turzo[a]

[a]Department of Computer Science & Engineering, Varendra University, Rajshahi, Bangladesh

*Corresponding author: kumarbiplob336@gmail.com

ABSTRACT

In Bangladesh, Ecommerce is flourishing day by day especially in the time of crisis the world is facing. There are many platforms available on these sites among which Daraz is the most successful marketplace. This online platform allowed people the ease to do shopping but a large number of reviews and comments made it difficult to opt for the best option. In this paper, the focus is on cataloguing the positive and negative reviews. For this purpose various classifiers were used by using Python. Data cleaning was done and after application of Term Frequency -Inverse Data Frequency with Principal Component analysis it was found that Ridge classifier performed best with more training time then other classifiers and depicted high accuracy. This classifier could help different businesses on different platforms to identify the positive and negative reviews and can provide customers with details about the quality of products.

Keywords: Adaboost classifier, Principal component analysis, Python, Ridge classifier, Term Frequency-Inverse Data Frequency

Advanced Research in Computer Engineering | ISBN: 978-81-948951-2-1
Book Series: Research Transcripts in Computer, Electrical and Electronics Engineering | Vol: 02

1. INTRODUCTION

Ecommerce is a short form of “electronic commerce’ which significantly elaborates any kind of exchange of currency for services online. Every business in ecommerce informs about the type of products they to their customers. There are three categories in which ecommerce businesses sell their products which are B2B, B2C and B2G. They can also be differentiated on the way they sell their services. These options are branded ecommerce stores, Ecommerce market places and conversational commerce. Popular ecommerce sites in Bangladesh are BD jobs, Clickbd, Bikroy, BoiMela, Rokomari, foodpanda, Daraz, Chaldal etc. Daraz is an online marketplace and logistics company which run in South Asian and Southeast Asian markets and was founded in 2012. It's one of the best ecommerce sites in Bangladesh. Ecommerce market in Bangladesh made a quantum jump in 2017 and developed at an incredible 70% from 2016. Ecommerce is at its peak in Bangladesh in this pandemic crisis. In this paper, a large number of Daraz product reviews were scraped and each review got labelled according to sentiment-based analysis that might be positive or negative. Total 900 reviews were amassed from Daraz and results in the building of a sentiment analyser which can be trained using data already present so that new reviews and comments can be catalogued. Presently, five ecommerce sites are growing rapidly in huge markets of Bangladesh. The principle goal of this study is to anticipate their sales and to find their customer satisfaction level.

2. LITERATURE REVIEW

Examination work was done which presents a structure for utilizing text digging for get-together client's input. This strategy is utilized to assemble the top credits related to a gathering of gadgets. An examination of properties of various gadgets and their positive and negative qualities referenced by clients were utilized to improve the group of people yet to come of items. [1].

For removal of uncertainty regarding purchases online reviews may help in clearing all doubts. A tale strategy was utilized that consolidated the Bass/Norton model and estimation examination while utilizing verifiable deals information and online audit information was created for anticipating item deals[2].

For text sentiment analysis feature extraction is one of the key method. The corresponding algorithms have important effects. A novel methodology was introduced to extract features. Summed up TF-IDF include vectors were gotten by the presentation of semantic comparability of equivalent words. The neighbourhood examples of highlight vectors were related to OPSM biclustering calculation [3].

In an indagation utilizing film audits, it was discovered that standard technician learning strategies certainly beats human delivered baselines. Three AI techniques were utilized called Naïve Bayes, greatest entropy order, and backing vector machines performed well on conventional Topic based classification [4].

Electronic business is getting famous because of an enormous number of item surveys. Assessment mining was utilized to catch client surveys. Furthermore, it isolated audits into abstract articulations and target articulations. An epic multi-dimensional model was proposed for assessment mining which incorporated client's attributes and their conclusion about any item [5].

No conclusion has been drawn yet on the multi-domains applicability in Chinese. Comparison was done for ten approaches of aspect opinion extractions on Chinese corpora from seven domains. Compared methods include TF-based models plus POS, CRFs based opinion mining, and SVM bas4d opinion mining CART based opinion mining and LPM-based opinion mining [6].

Term selection methods reduce the size of vocabulary effectively in text categorization for improving quality of classifier. It focuses on identification of relevant terms for each category without affecting the quality of text categorization [7].

An approach was described to object reveal which searches and localize all the occurrences of an object in a video. The object is represented by a set of view point invariant region descriptors so that recognition can proceed successfully despite changes in view point. Efficient retrieval can be achieved by employing methods from statistical text retrieval including inverted file system. In addition, weightings of text and document frequency. The final rankings also depend on spatial layout of the regions [8].

The joint likelihood capacity of groupings of words in a language can be educated as an objective of the factual language model. It is naturally troublesome. Successful yet conventional methodologies dependent on n-grams get speculation by connecting short covering arrangements found in the preparation set. To battle against the scourge of dimensionality a circulated portrayal for words was utilized which permitted each preparation succession to become familiar with the model about an outstanding number of semantically neighbouring sentences [9].

Variations of a neural organization design for factual language displaying have been proposed and applied effectively e.g in the language demonstrating part of the discourse recognizer. They get familiar with an implanting for words in a persistent space that assists with smoothing a language and give better speculation in any event, when the quantity of preparing models is deficient [10].

Two epic model structures for calculation of persistent vector portrayals of words from extremely enormous datasets were proposed. The quality was estimated in a word closeness task and the outcomes are contrasted with beforehand best-performing strategies dependent on various kinds of neural organizations [11].

With the quick development of internet shopping more clients share their encounters and item audits. Both huge amounts and different types of audits can carry trouble for likely customers to synopsis all heterogeneous surveys for reference. Another positioning framework was proposed through online audits dependent on different parts of elective items. Initially the loads of these angles were resolved with the LDA point model to figure the target notion estimation of the item. At that point purchaser's customized inclinations were taken. A coordinated diagram model was built and the last score was registered utilizing the PageRank calculation [12].

For processing natural language processing tasks continuous word representations were used. Popular methods which learn these representations ignore the wordings morphology. A vector representation is associated to each character n-grams. The method was fast which allowed to train models [13].

Archive feeling arrangement has become a zone of exploration. It tends to be viewed as a unique instance of effective characterization applied distinctly to emotional segments of an archive. Thus the critical errand in archive estimation order m is subjectivity. Approaches existing to remove emotionally depend on etymological assets [14].

Another troupe learning structure was utilized for assessment order of Chinese online surveys. Most importantly as indicated by convoluted attributes of Chinese online audits a grammatical form was separated. Calculation of Random Subspace dependent on data picked up by considering the issue of enormous highlights in the surveys which can improve base classifiers all the while [15].

These days clients search for highlights that help them explicitly yet among a great many audits, it's difficult to track down certain criticisms. The proposed framework followed a semantic-based way to deal with extricate highlights of items. The recursive profound model is utilized to distinguish the estimation direction of surveys [16].

The developing prevalence of conclusion rich assets, for example, survey gatherings for the items have made it hard to pick the correct item. A unique framework was proposed for highlight based synopsis of clients' sentiments for online items. The last extremity of highlight conclusion sets was determined [17].

A plan was made for Stanford composed conditions to give a basic portrayal of linguistic connections in a sentence that can without much of a stretch be perceived. As opposed to the expression structure portrayal that has since quite a while ago overwhelmed the computational phonetic network [18].

The web has become a great hotspot for social affair client audits. The quantity of surveys got by an item develops fastly. The nature of client audits is surveyed as the most critical. An endeavour was made to survey see dependent on its quality to assist them with settling on a legitimate choice [19].

In another research semantic based approach was used to assist customers and manufacture merchants to mine different products features. And to find opinion summarisation about each extracted features [20].

3. METHODOLOGY

All the data regarding reviews and comments were scraped the website of Daraz. A sum total of 900 comments were scraped from this website and among them 593 are 0's and 307 are 1's. 0 and 1 (positive or negative) was manually set after reading the whole sentence by human. The data set was then filtered and emoticons were removed. All comments are in bangla language.

The methodological steps followed are as follows:

- First comments were gathered from the relevant website.
- Specific sentences were picked from reviews or comments.
- Stop words were anticipated from each sentence.
- By utilizing TF-IDF text sentence data was converted into numerical data.
- Final dataset was obtained by applying PCA.
- A variety of machine learning algorithms were applied on datasets of Python (pycaret).
- At the end dataset was analyse.

The fig 1 is image representation of methodological steps.

At first, all non- English words and stop words were discarded. The positive comments were labelled as numeric 0 and negative comments got catalogued as numeric 1 (you can see it on fig 2). All the non-English words got axed from it by us. Natural Language Toolkit (NLTK) information center of python is being used for this purpose. Ergo, after the moping through the process, on the norm 1983 data set were collected.

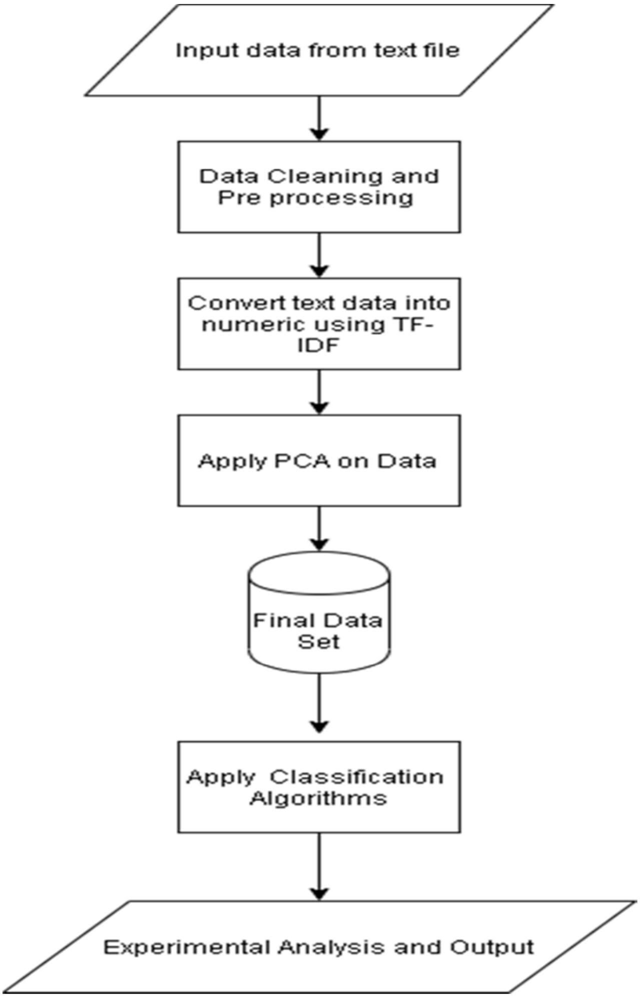
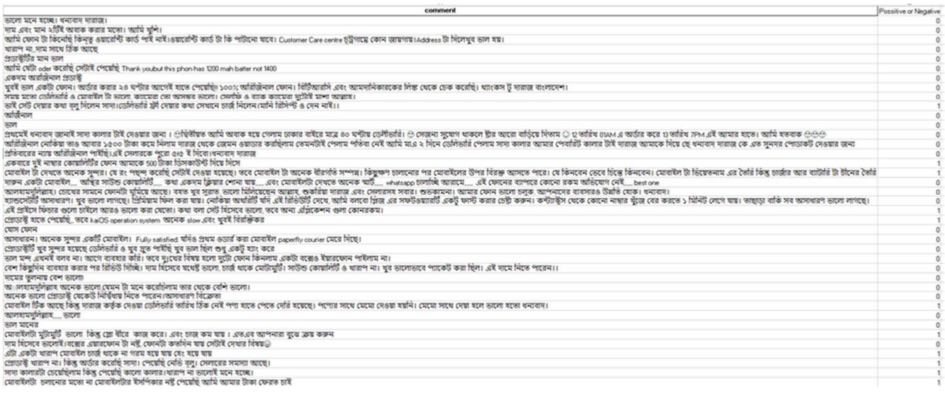

Input data from text file
Data Cleaning and Pre processing
Convert text data into numeric using TF-IDF
Apply PCA on Data
Final Data Set
Apply Classification Algorithms
Experimental Analysis and Output

Fig. 1. Vertical Workflow



comment	Possitive or Negative
ভালো মনে হচ্ছে। ধন্যবাদ দারাজ।	0
দাম এবং মান ২টিই অবাক করার মতো। আমি খুশি।	0
আমি ফোন টা কিনেছি কিন্তু ওয়ারেন্টি কার্ড পাই নাই।ওয়ারেন্টি কার্ড টা কি পাঠানো যাবে। Customer Care centre চট্টগ্রামে কোন জায়গায়।Address টা দিলেখুব ভাল হয়।	0
খারাপ না_দাম সাথে ঠিক আছে	0
প্রডাক্টটির মান ভাল	0
আমি যেটা oder করেছি সেটাই পেয়েছি Thank youbut this phon has 1200 mah batter not 1400	0
একদম অরজিনাল প্রডাক্ট	0
খুবই ভাল একটা ফোন। অর্ডার করার ২৪ ঘন্টার আগেই হাতে পেয়েছি। ১০০% অরিজিনাল ফোন। বিটিআরসি এবং আমদানিকারকের লিঙ্ক থেকে চেক করেছি। থ্যাংকস টু দারাজ বাংলাদেশ।	0
সময় মতো ডেলিভারি ও মোবাইল টা ভালো, ক্যামেরা তো অসম্ভব ভালো। সেলফি & ব্যাক ক্যামেরা দুটোই মাশা আল্লাহ।	0
তাই সেট দেয়ার কথা বলু দিলেন সাদা।ডেলিভারি ফ্রী দেয়ার কথা সেখানে চার্জ নিলেন।মানি রিসিপ্ট ও দেন নাই।।	1
অরিজিনাল	0
ভাল	0
প্রথমেই ধন্যবাদ জানাই সাদা কালার টাই দেওয়ার জন্য । 😊দ্বিতীয়ত আমি অবাক হয়ে গেলাম ঢাকার বাইরে মাত্র ৪০ ঘন্টায় ডেলিভারি। 😊 সেজন্য সুযোগ থাকলে ষ্টার আরো বাড়িয়ে দিতাম 😊 12 তারিখ 01AM এ অর্ডার করে 13 তারিখ 7PM এই আমার হাতে। আমি হতবাক 😊😊😊	0
অরিজিনাল নোকিয়া তাও আবার ১৫০০ টাকা কমে নিলাম দারাজ থেকে যেমন ওয়াডার করছিলাম তেমনটাই পেলাম পার্থক্য নেই আমি মাত্র ২ দিনে ডেলিভারি পেলাম সাদা কালার আমার ফেবারিট কালার টাই দারাজ আমাকে দিয়ে ছে ধন্যবাদ দারাজ কে এত সুন্দর পোডাক্ট দেওয়ার জন্য	0
প্রতিবারের ন্যায় অরিজিনাল পাইছি।এই সেলারকে পুরো ৫/৫ ই দিবো।ধন্যবাদ দারাজ	0
একবারে দুই নাম্বার কোয়ালিটির ফোন আমাকে 500 টাকা ডিসকাউন্ট দিয়ে দিসে	1
মোবাইল টা দেখতে অনেক সুন্দর। যে রং পছন্দ করেছি সেটাই দেওয়া হয়েছে। তবে মোবাইল টা অনেক ধীরগতি সম্পন্ন। কিছুক্ষণ চালানোর পর মোবাইলের উপর বিরক্ত আসতে পারে। যে কিনবেন ভেবে চিন্তে কিনবেন। মোবাইল টা ভিয়েতনাম এর তৈরি কিন্তু চার্জার আর ব্যাটারি টা চীনের তৈরি	1
দারুন একটা মোবাইল__ অস্থির সাউন্ড কোয়ালিটি__ কথা একদম ক্লিয়ার শোনা যায়__এবং মোবাইলটা দেখতে অনেক স্মার্ট__ whatsapp চালাচ্ছি আরামে__ এই ফোনের ব্যাপারে কোনো রকম অভিযোগ নেই__ best one	0
আলহামদুলিল্লাহ। চোখের সামনে ফোনটা ঘুমিয়ে আছে। বহত খুব সুরাত, ভালো মিলিয়েছেন আল্লাহ, শুকরিয়া দারাজ এবং সেলারসহ সবার। শুভকামনা। আমার ফোন ভালো চলুক, আপনাদের ব্যবসারও উন্নতি হোক। ধন্যবাদ।	0
হ্যান্ডসেটটি অসাধারণ। খুব ভালো লাগছে। প্রিমিয়াম ফিল করা যায়। নোকিয়া অথরিটি যদি এই রিভিউটি দেখে, আমি বলবো প্লিজ এর সফটওয়্যারটি একটু ফাস্ট করার চেষ্টা করুন। কন্টাক্টস থেকে কোনো নাম্বার খুঁজে বের করতে ১ মিনিট লেগে যায়। তাছাড়া বাকি সব অসাধারণ ভালো লাগছে।	0
এই প্রাইসে ফিচার গুলো চাইলে আরও ভালো করা যেতো। কথা বলা সেট হিসেবে ভালো, তবে অন্য এপ্লিকেশন গুলা কোনরকম।	1
প্রোডাক্ট হাতে পেয়েছি_ তবে kaiOS operation system অনেক slow এবং খুবই বিরক্তিকর	1
যোস ফোন	0
অসাধারন। অনেক সুন্দর একটি মোবাইল। Fully satisfied. যদিও প্রথম ওডার্র করা মোবাইল paperfly courier মেরে দিছে।	0
প্রোডাক্টটি খুব সুন্দর হয়েছে ডেলিভারি ও খুব দ্রুত পাইছি খুব ভাল ছিল শুধু একটু হ্যাং করে	0
ভাল মন্দ এখনই বলব না। আগে ব্যবহার করি। তবে দুঃখের বিষয় হলো দুটো ফোন কিনলাম একটা বক্সেও ইয়ারফোন পাইলাম না।	0
বেশ কিছুদিন ব্যবহার করার পর রিভিউ দিচ্ছি। দাম হিসেবে যথেষ্ট ভালো, চার্জ থাকে মোটামুটি। সাউন্ড কোয়ালিটি ও খারাপ না। খুব ভালোভাবে প্যাকেট করা ছিল। এই দামে নিতে পারেন।।	0
দামের তুলনায় বেশ ভালো।	0
আ◌লহামদুলিল্লাহ অনেক ভালো,যেমন টা মনে করেছিলাম তার থেকে বেশি ভালো।	0
অনেক ভালো প্রোডাক্ট যেকেউ নির্দ্বিধায় নিতে পারেন।অসাধারণ বিক্রেতা	0
মোবাইল ঠিক আছে কিন্তু দারাজ কর্তৃক দেওয়া ডেলিভারি তারিখ ঠিক নেই পণ্য হাতে পেতে দেরি হয়েছে। পণ্যের সাথে মেমো দেওয়া হয়নি। মেমো সাথে দেয়া হলে ভালো হতো ধন্যবাদ।	1
আলহামদুলিল্লাহ__ ভালো	0
ভাল মানের	0
মোবাইলটা মুটামুটি ভালো কিন্তু স্লো ধীরে কাজ করে। এবং চাজ কম যায় । এতএব আপনারা বুঝে ক্রয় করুন	1
দাম হিসেবে ভালোই।বক্সের এয়ারফোন টা নষ্ট, ফোনটা কতদিন যায় সেটাই দেখার বিষয়😊	0
এটা একটা খারাপ মোবাইল চার্জ থাকে না গরম হয়ে যায় হেং হয়ে যায়	1
প্রোডাক্ট খারাপ না। কিন্তু অর্ডার করেছি সাদা। পেয়েছি নেভি বলু। সেলারের সমস্যা আছে।	1
সাদা কালারটা চেয়েছিলাম কিন্তু পেয়েছি কালো কালার।খারাপ না ভালোই মনে হচ্ছে।	1
মোবাইলটা চলানোর মতো না মোবাইলটার ইয়ার্পিকার নষ্ট পেয়েছি আমি আমার টাকা ফেরত চাই	1

Fig. 2. Fig 2: Data before TH-IDF

An analytical statistic is a numerical or scientific form of statistic which is being contemplated to mirror the principal of word in a docket or corpus and is called Short Term Frequency-Inverse Document Frequency (TF-IDF). This factor has weightage in retrieving information, text mining and user modelling through hunting of this data.

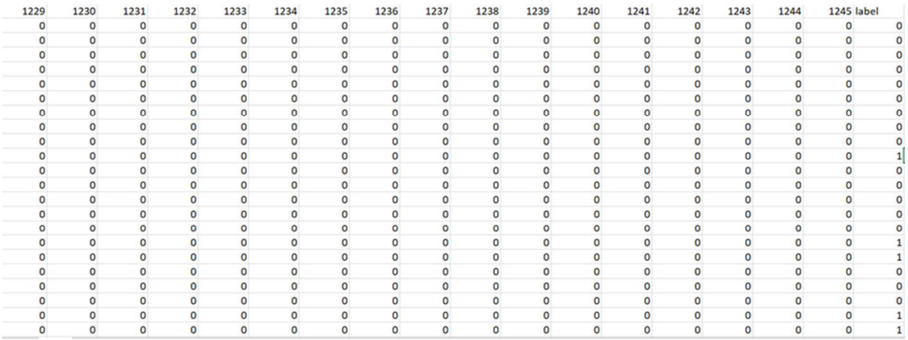
Frequency of a word which pops up in a docket divided by the gross number of words in the document. Every document has its own term frequency.

The log of the documents number divided by word w containing documents. Inverse data frequency determines the weight of rare words across all documents in the corpus.

Most work is being done from Scikit-Learn. Text data is taken by it and converted to numeric information set. After this conversion, our data has 3394 features. Due to so many less important features, features extraction were done using PCA.

1. Principal Component Analysis

A new coordinate system is being metamorphosed from data through orthogonal linear transformation so that each coordinate has greatest variance by scalar projection of data in an ordered way and so on. This is called principal component analysis. Higher variance comes to lie in first coordinate which is called first principal component and the lower variance in second coordinate. Our information set has 1678 features after application of principal component analysis. When applications of dimensions of principal component analysis got reduced and the data quality got lost. In case of principal quality analysis, 95% calibre of data was being maintained. 95% of the quality of real data was preserved by setting value of 'n' components as 0.95. Our latest data has 1678 characteristics after application of principal component analysis.



1229	1230	1231	1232	1233	1234	1235	1236	1237	1238	1239	1240	1241	1242	1243	1244	1245	label
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Fig. 3. Data after TF-IDF

There are 1246 different fields (fig 3) of numeric data in the processed data in which last field signifies 1 for negative comments and 0 for positive comments.

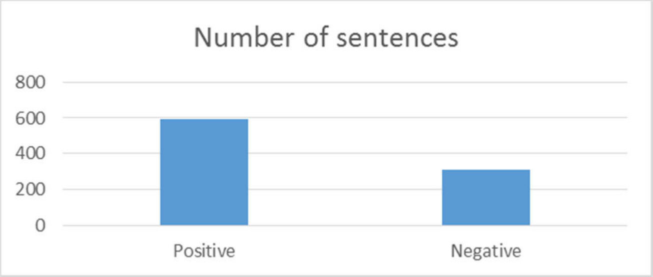


Fig. 4. Number of Positive and Negative reviews

4. EXPERIMENTAL RESULTS

Following are the result obtained after implementation of the dataset in MATLAB: Among 15 classifiers accuracy, prediction speed, training time and total misclassification of top four were used.

Table 1. Summery of Top 4 Classifiers performance

Classifier Name	**Accuracy**	**AUC**	**Recall**	**Precision**	**F1**	**Kappa**	**Time (sec)**
SVM (Linear)	0.7631	0.0000	0.5050	0.7212	0.5919	0.4323	0.1742
Extra Trees Classifier	0.7615	0.7957	0.4305	0.7856	0.5487	0.4057	0.7022
Ridge Classifier	0.7520	0.0000	0.4496	0.7283	0.5524	0.3939	0.0958
AdaBoost Classifier	0.7472	0.7535	0.5149	0.6714	0.5805	0.4045	0.7662

Ridge classifier has the fastest training speed while subspace Adaboost has the slowest as depicted in the training time graph. The Naïve Bayes decision tree was discarded due to low precision.

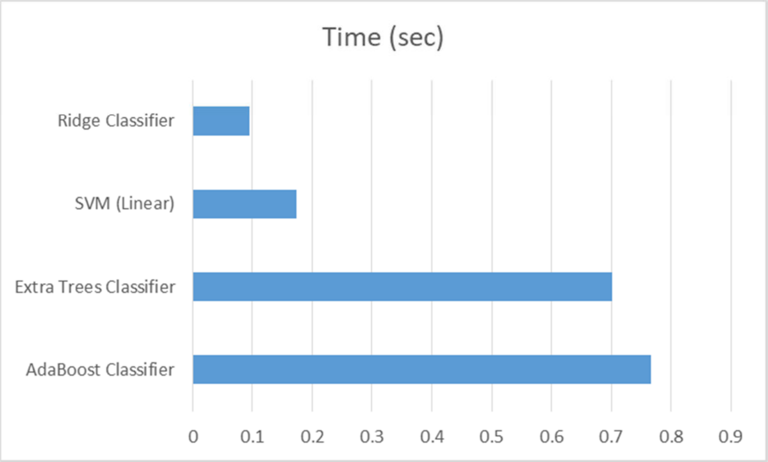


Fig. 5. Time comparison of top 4 performers

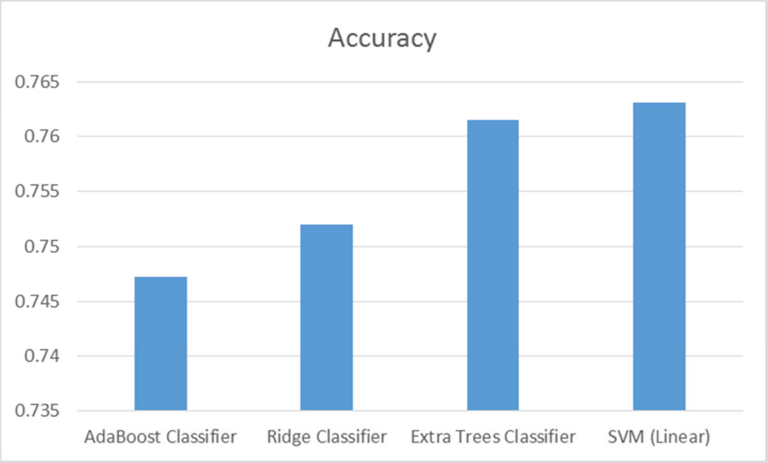


Fig. 6. Accuracy comparison of top 4 performers

Linear SVM gave the highest accuracy followed by Extra Trees Classifier.

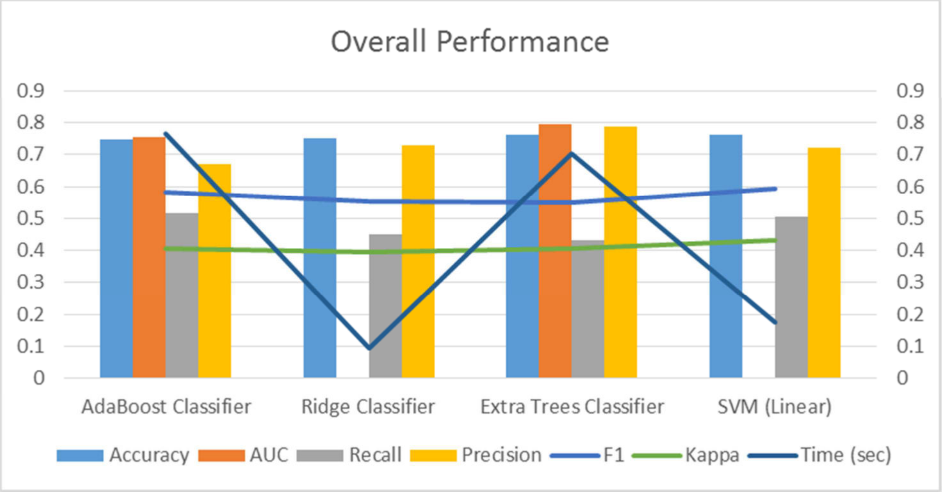
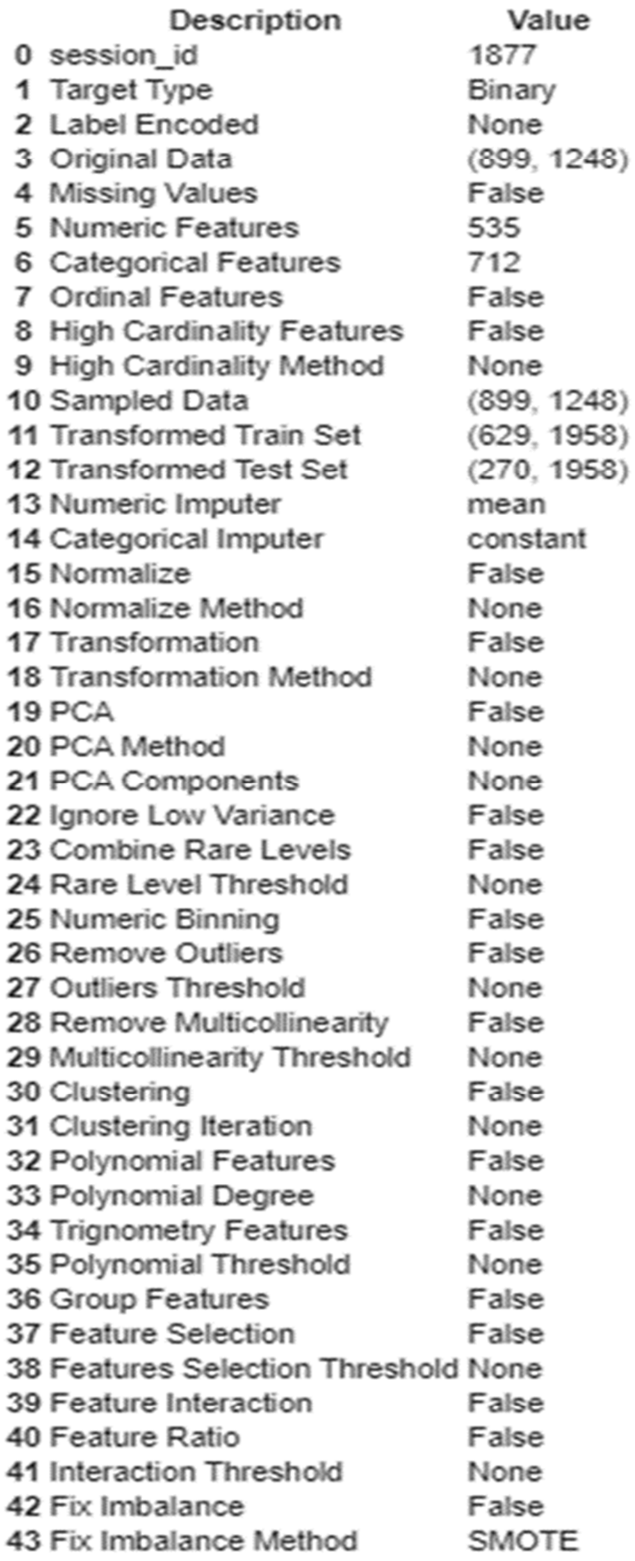


Fig. 7. Overall Performance of top 5 classifiers

Though Ridge classifier has performed best but has low training speed. Linear SVM has high precision with high anticipation speed. So in case of implementation of Python for optimization ridge classifier has done the best function for categorizing positive and negative comments or reviews of Bangla.

Here are the preset values used for the classifiers-



	Description	Value
0	session_id	1877
1	Target Type	Binary
2	Label Encoded	None
3	Original Data	(899, 1248)
4	Missing Values	False
5	Numeric Features	535
6	Categorical Features	712
7	Ordinal Features	False
8	High Cardinality Features	False
9	High Cardinality Method	None
10	Sampled Data	(899, 1248)
11	Transformed Train Set	(629, 1958)
12	Transformed Test Set	(270, 1958)
13	Numeric Imputer	mean
14	Categorical Imputer	constant
15	Normalize	False
16	Normalize Method	None
17	Transformation	False
18	Transformation Method	None
19	PCA	False
20	PCA Method	None
21	PCA Components	None
22	Ignore Low Variance	False
23	Combine Rare Levels	False
24	Rare Level Threshold	None
25	Numeric Binning	False
26	Remove Outliers	False
27	Outliers Threshold	None
28	Remove Multicollinearity	False
29	Multicollinearity Threshold	None
30	Clustering	False
31	Clustering Iteration	None
32	Polynomial Features	False
33	Polynomial Degree	None
34	Trignometry Features	False
35	Polynomial Threshold	None
36	Group Features	False
37	Feature Selection	False
38	Features Selection Threshold	None
39	Feature Interaction	False
40	Feature Ratio	False
41	Interaction Threshold	None
42	Fix Imbalance	False
43	Fix Imbalance Method	SMOTE

Fig. 8. Preset values of the classifiers

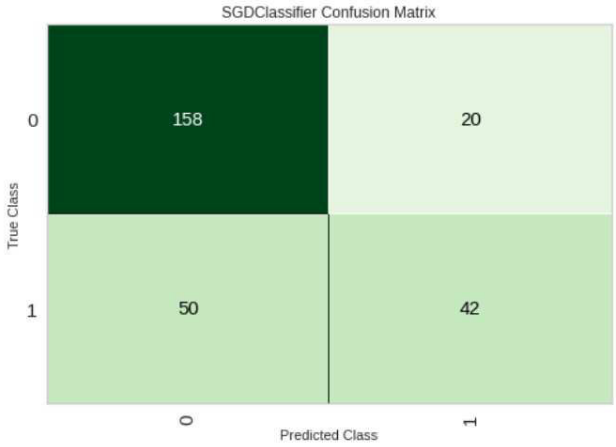


Fig. 9. Confusion Matrix of Linear SVM

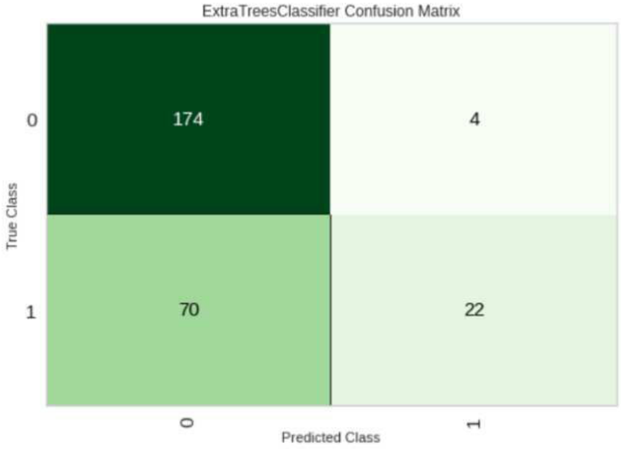


Fig. 10. Confusion Matrix of Extra Trees Classifier

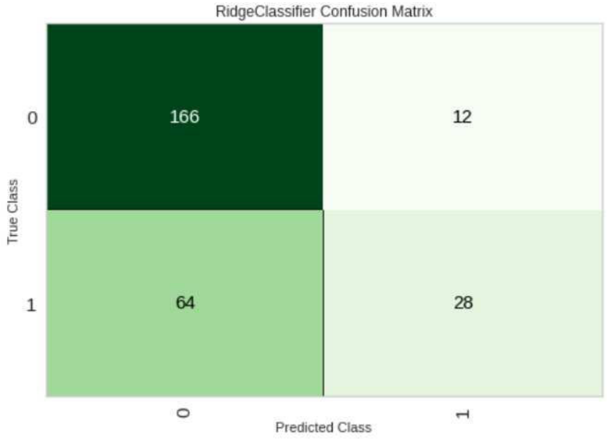


Fig. 11. Confusion Matrix of Ridge Classifier

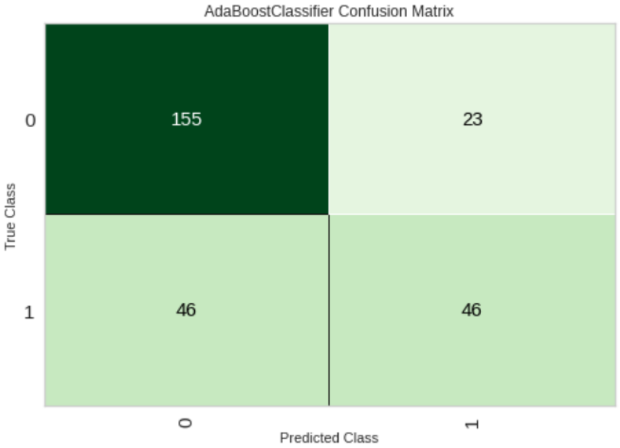


Fig. 12. Confusion Matrix of Adaboost Classifier

From the confusion matrix, its clarified Ridge classifier has high value as well as for extra trees classifier. ROC curve of various classifiers are given in Fig 13.

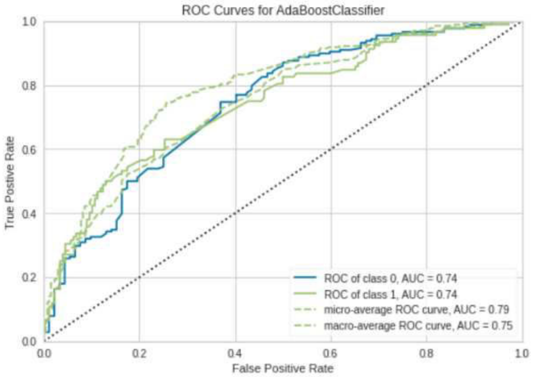


Fig. 13. Adaboost Classifier

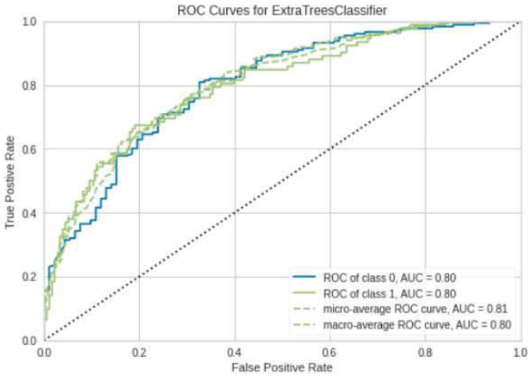
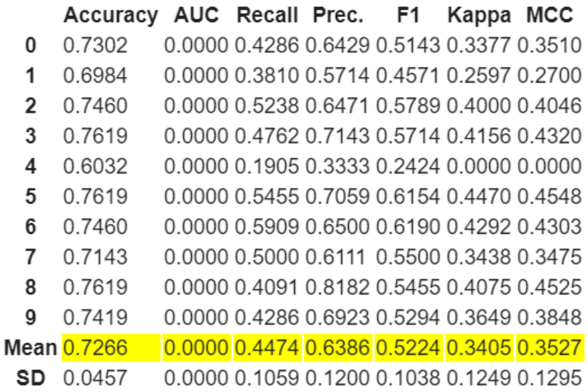


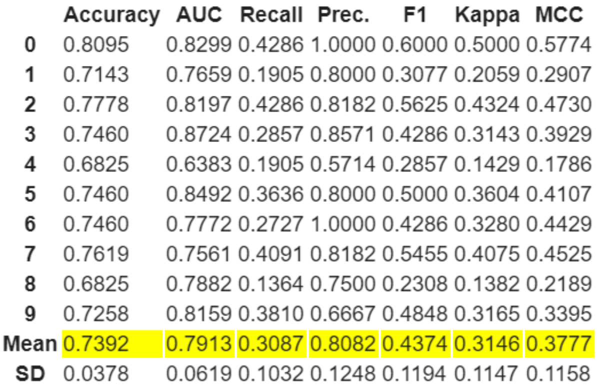
Fig. 14. Extra Trees Classifier

For ROC curves a steeper curve depicts better output and Extra trees classifier has steeper curves.



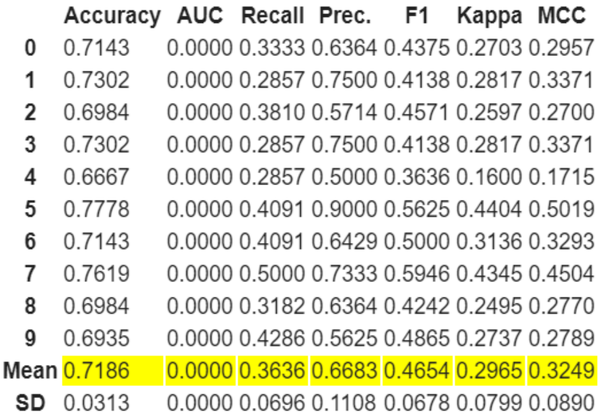
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7302	0.0000	0.4286	0.6429	0.5143	0.3377	0.3510
1	0.6984	0.0000	0.3810	0.5714	0.4571	0.2597	0.2700
2	0.7460	0.0000	0.5238	0.6471	0.5789	0.4000	0.4046
3	0.7619	0.0000	0.4762	0.7143	0.5714	0.4156	0.4320
4	0.6032	0.0000	0.1905	0.3333	0.2424	0.0000	0.0000
5	0.7619	0.0000	0.5455	0.7059	0.6154	0.4470	0.4548
6	0.7460	0.0000	0.5909	0.6500	0.6190	0.4292	0.4303
7	0.7143	0.0000	0.5000	0.6111	0.5500	0.3438	0.3475
8	0.7619	0.0000	0.4091	0.8182	0.5455	0.4075	0.4525
9	0.7419	0.0000	0.4286	0.6923	0.5294	0.3649	0.3848
Mean	0.7266	0.0000	0.4474	0.6386	0.5224	0.3405	0.3527
SD	0.0457	0.0000	0.1059	0.1200	0.1038	0.1249	0.1295

Fig. 15. Linear SVM



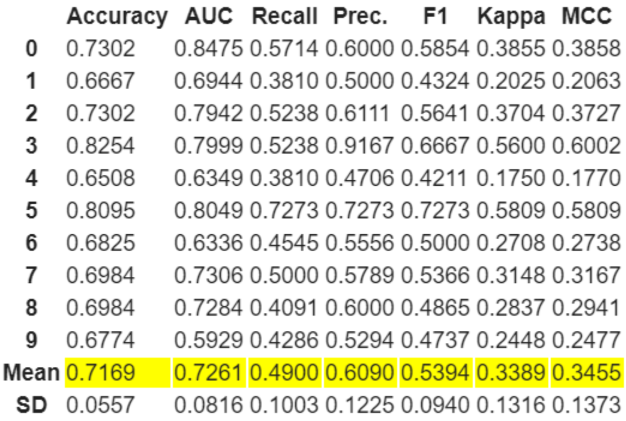
	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.8095	0.8299	0.4286	1.0000	0.6000	0.5000	0.5774
1	0.7143	0.7659	0.1905	0.8000	0.3077	0.2059	0.2907
2	0.7778	0.8197	0.4286	0.8182	0.5625	0.4324	0.4730
3	0.7460	0.8724	0.2857	0.8571	0.4286	0.3143	0.3929
4	0.6825	0.6383	0.1905	0.5714	0.2857	0.1429	0.1786
5	0.7460	0.8492	0.3636	0.8000	0.5000	0.3604	0.4107
6	0.7460	0.7772	0.2727	1.0000	0.4286	0.3280	0.4429
7	0.7619	0.7561	0.4091	0.8182	0.5455	0.4075	0.4525
8	0.6825	0.7882	0.1364	0.7500	0.2308	0.1382	0.2189
9	0.7258	0.8159	0.3810	0.6667	0.4848	0.3165	0.3395
Mean	0.7392	0.7913	0.3087	0.8082	0.4374	0.3146	0.3777
SD	0.0378	0.0619	0.1032	0.1248	0.1194	0.1147	0.1158

Fig. 16. Extra Trees Classifier



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7143	0.0000	0.3333	0.6364	0.4375	0.2703	0.2957
1	0.7302	0.0000	0.2857	0.7500	0.4138	0.2817	0.3371
2	0.6984	0.0000	0.3810	0.5714	0.4571	0.2597	0.2700
3	0.7302	0.0000	0.2857	0.7500	0.4138	0.2817	0.3371
4	0.6667	0.0000	0.2857	0.5000	0.3636	0.1600	0.1715
5	0.7778	0.0000	0.4091	0.9000	0.5625	0.4404	0.5019
6	0.7143	0.0000	0.4091	0.6429	0.5000	0.3136	0.3293
7	0.7619	0.0000	0.5000	0.7333	0.5946	0.4345	0.4504
8	0.6984	0.0000	0.3182	0.6364	0.4242	0.2495	0.2770
9	0.6935	0.0000	0.4286	0.5625	0.4865	0.2737	0.2789
Mean	0.7186	0.0000	0.3636	0.6683	0.4654	0.2965	0.3249
SD	0.0313	0.0000	0.0696	0.1108	0.0678	0.0799	0.0890

Fig. 17. Ridge Classifier



	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7302	0.8475	0.5714	0.6000	0.5854	0.3855	0.3858
1	0.6667	0.6944	0.3810	0.5000	0.4324	0.2025	0.2063
2	0.7302	0.7942	0.5238	0.6111	0.5641	0.3704	0.3727
3	0.8254	0.7999	0.5238	0.9167	0.6667	0.5600	0.6002
4	0.6508	0.6349	0.3810	0.4706	0.4211	0.1750	0.1770
5	0.8095	0.8049	0.7273	0.7273	0.7273	0.5809	0.5809
6	0.6825	0.6336	0.4545	0.5556	0.5000	0.2708	0.2738
7	0.6984	0.7306	0.5000	0.5789	0.5366	0.3148	0.3167
8	0.6984	0.7284	0.4091	0.6000	0.4865	0.2837	0.2941
9	0.6774	0.5929	0.4286	0.5294	0.4737	0.2448	0.2477
Mean	0.7169	0.7261	0.4900	0.6090	0.5394	0.3389	0.3455
SD	0.0557	0.0816	0.1003	0.1225	0.0940	0.1316	0.1373

Fig. 18. Adaboost Classifier

The tuning of results showed that having low training time as in the case of Ridge classifier its variant can give 77.78% accuracy and this is the reason Ridge classifier is the best.

5. CONCLUSION

As the results depicted that after gathering all data different algorithms were applied among which Ridge classifier has the fast training time than subspace Adaboost while Naïve Bayes was discarded due to its inferior performance. Moreover the tuning of these classifiers also inferred that regardless of the training time Ridge classifier can give the precision of up to 77.8% and this makes it the best classifier to be used for categorizing comments or reviews in any business.

REFERENCES

1. L. J. a. Y. Tsai, "Using Text Mining of Amazon Reviews to Explore User-Defined Product Highlights and Issues”, in Int'l Conf. Data Mining , CA,USA, 2015.
2. Y.-J. Zhi-Ping, "Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis”, Journal of Business Research, pp. 90-100, 2017.
3. 3. ·. Y. X. ·. H. Z. ·. X. L. ·. X. H. ·. Z. M. Xin Chen1, "A novel feature extraction methodology for sentiment analysis of product reviews”, Neural Computing and Applications, 2018.
4. B. P. a. L. Lee, "Thumbs up? Sentiment Classification using Machine Learning Techniques”, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, 2002.

5. l. R. Y. Y. L. Zhang, "Integration of Sentiment Analysis into Customer Relational Model: The Importance of Feature Ontology and Synonym", Procedia Technology, vol. 11, pp. 495-501, 2013 .
6. G. T. &. H. W. Wei Wang, "Cross-domain comparison of algorithm performance in extracting aspect-based opinions from Chinese online reviews", International Journal of Machine Learning and Cybernetics, pp. 1053-1070, 2016.
7. T. B. &. C. A. Murthy, "A supervised term selection technique for effective text categorization", International Journal of Machine Learning and Cybernetics, vol. 7, pp. 877-892, 2015.
8. J. Sivic and A. Zisserman, "Efficient Visual Search of Videos Cast as Text Retrieval", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 4, pp. 591-606, April 2009, doi: 10.1109/TPAMI.2008.111.
9. Y. Bengio, "A Neural Probabilistic Language Model", Journal of Machine Learning Research, Vol. 3, 2003.
10. F. Morin, "Hierarchical Probabilistic Neural Network Language Model", in Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, 2005.
11. K. C. G. C. J. D. Tomas Mikolov, "Efficient Estimation of Word Representations in Vector Space", Cornell university, vol. 3, 2013.
12. C. G. Z. D. X. Kou, "Products Ranking Through Aspect Based Sentiment Analysis Of Online Heterogeneous Reviews", vol. 5, no. 27, pp. 542-558, 2018.
13. E. G. A. J. T. M. Piotr Bojanowski, "Enriching Word Vectors with Subword Information", 2017.
14. K. Sarvabhotla, "Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents", Information Retrieval, vol. 14, pp. 337-353, 2011.
15. Y. X. Jiafeng Huang, "Sentiment analysis of Chinese online reviews using ensemble learning framework", Cluster Computing, vol. 22, pp. 3043-3058, 2018.
16. N. Devasia and R. Sheik, "Feature extracted sentiment analysis of customer product reviews", International Conference on Emerging Technological Trends (ICETT), Kollam, 2016, pp. 1-6, doi: 10.1109/ICETT.2016.7873646.
17. K. B. DurgaToshniwal, "Feature based Summarization of Customers' Reviews of Online Products", Procedia Computer Science, vol. 22, pp. 142-151, 2013.
18. M.-C. d. M. a. C. D. Manning, "Stanford typed dependencies manual", 2008.

19. S. S. Prakash Hiremath, "Cluster Analysis of Customer Reviews Extracted from Web Pages", Journal of Applied Computer Science & Mathematics, vol. 9, no. 4, 2010.
20. R. k. V. a. K. Raghuveer, "Web User Opinion Analysis for Product Features Extraction and Opinion Summarization", International Journal of Web & Semantic Technology (IJWesT), vol. 3, no. 4, 2012.

Cite this article

Biplob Kumar, Pritom Sarker and Nakib Aman Turzo, Application of Classifiers for Assortment of Online Reviews, In: Sandip A. Kale editor, Advanced Research in Computer Engineering, Pune: Grinrey Publications, 2021, pp. 67-82