NUML International Journal of Engineering and Computing
Volume: 4 Issue 1
https://numl.edu.pk/journals/nijec
Print ISSN: 2788-9629
E-ISSN:2791-3465
DOI:https://doi.org/10.52015/nijec.v4i1.90

An Information Processing Model for Assessment of User Reviews

Hamid Nawaz[a], Naila Batool[a], Muhammad Tahir[b], Muhammad Usman[c]

[a] Department of Software Engineering, Govt. College University, Faisalabad, Pakistan
[b] School of Electronics and Information Engineering, Changchun University of Science and Technology, China
[c] Department of Computer Science, Riphah International University, Faisalabad

Submitted	Revised	Published
02-Aug-2025	20-Oct-2025	31-Oct-2025

Abstract

This Analysis of reviews became a valuable source of accurate information. In this research we will analyze reviews. Analyzed reviews will create an accurate data set. Fitting an appropriate model is necessary to get accuracy. Problems in fitting an appropriate model are under-fitting and over-fitting. An under-fit model will be less flexible and cannot account for the data. Over-fitting is a modeling error that occurs when a function is too closely fit to a limited set of data points. To solve the above mentioned problems, features and appropriate algorithms are selected. As a solution we will perform preprocessing in a better way along with the machine learning algorithms to assess the impact of preprocessing. The main focus of this research is to assess the impact of preprocessing steps on different classifiers.

Keywords: text preprocessing, user reviews, over-fitting, natural language processing

Corresponding author: hamidnawaz844@gmail.com

1. Introduction

The emerging and rapid growth of technology has created a bulk of data. Majorly, social networking websites are a great source of data. It has become a big challenge for data scientists to mine this data and find precise and accurate information. The reviews could be in variety. Individuals of different ages, from different parts of the world, utilize online media to communicate their feelings concerning various subjects, such as occasions, administrations, and items. These reviews could be a foundation of precious feedback for individuals and administrations; this feedback permits them to develop new and better products and services. The study of these reviews could partially substitute traditional review polls, which are typically slow and expensive. Social networking websites produce data in the form of reviews regarding any product, service, movie, or any other entity. Because of this big data, there is a need for some fruitful techniques to make this useful. So preprocessing of reviews or text is the most appropriate way to achieve the goal for data scientists. The reviews identified

correctly can become a road map for decision making. Text information contains characters, similar to punctuations, stop words, etc., which does not give data and increment the examination's intricacy. To rearrange our information, we eliminate this noise to acquire a perfect and analyzable dataset. Preprocessing of text intends to acquire the content in a structure which is an unsurprising and analyzable particular undertaking. That particular task identifies with a particular space. An overall system for moving toward the textual information tasks comprises five stages: a collection of data, preprocessing of text, exploration and perception of the text, manufacturing the necessary model, and the last step is assessing the model. This system works in a loop format by nature.

Text preprocessing is customarily a significant advance for natural language processing (NLP) assignments. It changes text into a more absorbable structure so that AI calculations can perform better. Text preprocessing procedures have various stages of changing over the content into the required arrangement. The essential structure of text preprocessing incorporates Tokenization, division, standardization, noise evacuation, etc.

Analysis of reviews is essential for preprocessing. Preprocessing reviews means making the reviews usable in decision making by reducing the data that is of no use. Better reviews will generate better results. This research aims to analyze the reviews and evaluate the accuracy of different machine learning algorithms by applying different preprocessing steps on reviews. For this purpose fitting, an appropriate model is necessary. Without fitting an appropriate model, it is not possible to get the desired results.

Significant problems in model fitting are under-fitting and over-fitting. Under-fitting happens when a proposed framework is not complex enough to capture dealings among a dataset's features and labels accurately [22]. Over-fitting happens when a model becomes too familiar with the dataset on which it trains; therefore, it loses its relevance to some other dataset. A framework is over-fitted when it is so specific to the original data that applying it to information gathered later on would bring about testing or off base results and less than the best outcome.

To solve problems, feature selection is applied. In this technique, features and appropriate algorithms are selected. In this research, another technique applies, which is called preprocessing. In this research, the central focus is on the preprocessing of reviews. The reviews are collected from Twitter. The data set on which preprocessing is done will generate a better result and increase the accuracy of the machine learning algorithm. Here we will not change the data set nor add anything extra to the existing data set. The main focus is on the preprocessing of reviews which are used in the data set. If the preprocessing is performed in an improved way, we can enhance the performance of the algorithm. In some cases, the performance of the algorithm increases. The advancement of technology is producing the bulk of data in a single second. This whole data is not useful for any single user. Most of the data is useless, which consumes extra memory and uses other system resources, which affects the system's cost and efficiency. To extract meaningful data from a large amount of data, preprocessing of text is essential. In what fields of work where prediction is essential to precede the work, preprocessing plays a significant role. With the help of text pre-processing, the bulk of data minimizes a usable format with a proper data form. The preprocessed data does not contain any extra and irrelevant material

2. Literature Review

[1] Intensely focuses on the self-explanation of the machine learning model. In this research, the author concluded that preprocessing is an essential step. Preprocessing affects the predictions very well. Many decisions that affect predictive behavior are taken at the preprocessing step. These steps and decisions consider the essential steps for prediction. These steps perform on two datasets, i.e., Students' Academic Performance and German Credit. The primary purpose is to identify the predictors for volatility that does not depend upon the preprocessing step and datasets. The decision tree helped the author to detect which features of more worth to inspect. It concludes that volatility is the necessary thing that helps in dual dimension, i.e., making predictions better and understanding the importance of preprocessing.

Smart Processing for Streaming Dataset (SPSD) handles model training and preprocessing together. Smart Processing for Streaming Dataset (SPSD) divides normalization of every numeric feature from model training. This research's primary objective is to minimize the new models needed in-stream mining without compromising the quality. It is the main aim of preprocessing because we minimize the data while keeping the preprocessing quality maintained. The results showed that SPSD maintained a quality of 50 % by renormalizing the data without constructing new classification models. Compared with the outdated framework, these can benefit from about 30% to 50% from SPSD. It will eradicate new training prices and diminish the available full amount of models. First, to achieve acceptable accuracy, three smart processing parameters for streaming dataset, chunk size, matrix 1and 2 thresholds [2].

In the research [3] the effect of preprocessing on email detection is discussed. Spam email detection has become one of the most exciting research topics. The spam detection or classifier will work adequately when applied to the dataset that properly preprocesses. The steps used in this research are noise removal, stop word removal, stemming, lemmatization, and term frequency. What is the effect of preprocessing on the spam email detection algorithm discussed in this research? The experiment applied to two famous machine learning classifiers: Naïve Bayes and Support Vector Machine. The dataset consists of a total of 962 spam messages text messages which are publically available. These messages use accuracy as an evaluation matrix. Based on these results, the accuracy of spam detection improved.

Here [4], the issues happening with preprocessing and the effect of preprocessing on the accuracy are discussed. All the preprocessing and some other operations are performed on query data. The data set used in this research is the farmer query dataset. This dataset is in textual form and arranged in tabular form. Its main target is to retrieve meaningful information from the data. Preprocessing techniques are applied to reduce the size of the data and increase accuracy. Preprocessing text data results show that it is free from null values, errors, ambiguities, and any other unnecessary material. This research gives knowledge about data mining and in-depth information about data preprocessing and the effect of preprocessing on machine learning classifiers' accuracy. Preprocessing operations performed using Python language and its supporting libraries.

Another researcher in [5] describes a system that uses data preprocessing activities that include Feature Selection and Discretization. Feature selection and dimension reduction are common data mining approaches in large datasets. Here the high data dimensionality of the dataset due to its extensive feature set poses a significant challenge. In Preprocessing with the assistance of Feature selection algorithms, the different required features choose, these exercises improve the classifier's accuracy. After this progression, different classifiers utilize, for

example, Naive Bayes, Hidden Naive Bayes, and NBTree. The benefit of Hidden Naive Bayes is a data mining model that loosens up the Naive Bayes Method's contingent independent assumption.

The research [6] shows that the author's primary focus is upon Machine Learning with preprocessing as its mandatory and essential part. Data preprocessing is an essential step of machine learning to improve the prediction accuracy of machine learning algorithms. The classifiers use the preprocessed data for prediction purposes. The goal of this research is to detect type II diabetes mellitus (T2DM) early. Different preprocessing methods are applied. The accuracy of methods without applying the preprocessing steps is 75%, 67%, 65%, and 74% against Logistic Regression, Artificial Neural Network, Support Vector Machine, and Random Forest. After applying the same steps on LUDB2, calculated accuracies are 98%, 94%, 74%, and 100%. This variation is that this dataset contains no missing value, no noise, and no balance. However, after applying preprocessing, the accuracy of these datasets is improved significantly.

In [7], this research the impact of simple text preprocessing decisions on the Neural text classifiers. The text preprocessing decisions discussed in this paper are Tokenization, lemmatization, multiword grouping, and lowercasing. The main focus is on being careful and consistent when doing preprocessing and comparing the system. It concluded that simple tokenization impacts more than other preprocessing techniques like lemmatization. There are also cases when the dataset is domain-specific, and only Tokenization gives poor results. There are a variety of results based on the choice of dataset. This research is considered as the basis for learning the effect of preprocessing in depth. It would also help the analysis in deep learning. It will also help the researchers in carefully performing preprocessing decisions and comparing and evaluating different models.

In [8], the author discussed that data preprocessing is vital in data mining and other fields. It is the factor that decreases the cost and minimizes the data by removing all unnecessary material from the target data. In this research, data was cleaned by two machine learning techniques, namely Conditional Random Fields and Hidden Markov Model, with a semi-automatic preprocessing framework. Besides preprocessing, other techniques already apply, but they were not crucial in cost and increased preprocessing time. Data of any size can be dealt with the help of this proposed system or framework. This experiment was applied to Pakistan Telecommunication Company for the sake of training and testing. Some data is used for learning and this data is the output of preprocessing, and the other data used for testing. This proposed hybrid technique gave an accuracy of 95.90%, much more than some other techniques. The developed model can clean any type of data, either large or short. Especially the proposed model is an expert in cleaning addresses.

What is the impact of preprocessing on spam detection? It is discussed in [9]. Online surveys become an essential wellspring of data that demonstrates the general opinion about items and services, which influence the client's choice to buy an item or administration. Since not every single online review and remark is genuine, it is essential to recognize fake reviews. Many Machine learning procedures could be applied to recognize spam reviews by extricating valuable features from the survey's content utilizing Natural Language Processing (NLP). Numerous sorts of features could be utilized right now as linguistic features, Word Count, and n-gram include sets and pronouns. To extract features, many sorts of preprocessing steps perform before applying the classification technique, which may incorporate POS tagging, n-gram term frequencies, stemming, stop word and punctuation marks filtering, and so on. These preprocessing steps may influence the overall accuracy of the review of spam detection. Here research the impacts of preprocessing steps on the accuracy of reviews spam detection.

Many Machine Learning techniques are applied, for example, Support Vector Machine (SVM) and Naïve Bayes (NB), and a labeled dataset of Hotels reviews will be applied.

Effect of preprocessing on tweets not investigated in [10]. One of the remarkable misuses of social media is the use of social media bots for illegal purposes. Using such agents, many individuals, different organizations, and many institutions perform illegal activities like illegal recruitment, use people for the stock market, spread misinformation, make illegal trade, and these bots perform many other activities. The detection of such bots is most important because avoiding people from these harmful uses is also of great importance. Performing preprocessing and feature selection with machine learning algorithms is the central area of this research. From tweets and profiles, the Twitter accounts obtain; this research work is so sharp to differentiate between the original and the bot accounts. It is the aim of this research.

Here [11] diabetes mellitus is one of the most widely recognized maladies among individuals of all age gatherings, influencing youngsters, teenagers, and youthful grown-ups. There is an expanding enthusiasm for utilizing Machine learning methods to analyze these persistent illnesses. There is an expanding enthusiasm for utilizing Machine learning methods to analyze these persistent illnesses. Be that as it may, most clinical datasets' low quality restrains the development of proficient models for a forecast of diabetes mellitus. Without proficient preprocessing strategies, managing these sorts of informational collections prompts uncertain outcomes. This paper introduces a productive preprocessing system, including a mix of missing worth substitution and attribute subset choice techniques on a notable diabetes mellitus informational index. The outcomes show that the proposed procedure can improve applied classifiers' performance and outflanks the customary strategies regarding accuracy and precision in diabetes mellitus prediction.

Pre-processing is the most crucial step in data mining. In [12], focuses on improving the prediction accuracy of text data with preprocessing. For this purpose, the decision tree is the most crucial technique. We contrasted the outcomes and the J48 without discretization. The results show that the precision of J48 after discretization is better than J48 before discretization. It included a preprocessing stage while training, which does noise removal, removes some regular features, and helps improve the accuracy of email classification. With the training result, we accomplished a moderate expectation while experiencing another approaching email. Then again, we did not preprocess the dataset and get the output. Comparing both outputs appears to have improved exactness, and bogus positives were significantly decreased by 25.39% for the preprocessed dataset. In this manner, the test outcomes show that joining Naïve Bayes grouping with the best possible data preprocessing can improve the expectation exactness and demonstrate. The preprocessing stage has a more significant effect in executing the Naïve Bayes classifier, particularly with the reduced number of bogus positives.

The researcher [13] evaluated preprocessing techniques for text classification. Multiple forms of texts changed into a single form with the help of preprocessing tools. This paper gives many evaluation tools for English text classification. In this paper, preprocessing steps are removing stop words, the Tokenization, and stemming. It compares two different methods TF-IDF and chi-square, with the cosine similarity method. It is used for text preprocessing. Based on the evaluation matrix, in ten different categories, it gave better performance. The results obtained from this research showed that the preprocessing improves the system performance for text classification and also performs better text recognition.

[14] Applying natural language modeling strategies to new corpora expects users to convert records to information utilizing different preprocessing operations. In any case, the impacts of these changes are still inadequately comprehended. We depict a few contemplates that measure the effect of preprocessing in various structures, concentrating on point modeling applications. We find that numerous regular practices either have no quantifiable impact or negatively impact the wake of representing predispositions instigated by displaying determination. Finally, we give suggestions regarding preprocess content for beginner users of theme models looking to research their text corpora.

In the research [15], an enormous extent of online reviews present in open areas usually is useful. To a considerable extent it is poisonous. The reviews contain grammatical mistakes that build the quantity of features complex, making the ML model hard to prepare—considering how the information researchers spend around 80% of their time gathering, cleaning, and arranging their information. We investigated how much exertion should put resources into the preprocessing of crude reviews previously taking care of it to the best in class order models. With the assistance of four models on Jigsaw harmful comment order data, we showed that the model's preparation with no change produces a generally conventional model. Sometimes, applying even fundamental changes leads to awful execution and ought to be applied with an alert.

In the research [16], A significant advance in a Sentiment Analysis framework for text mining is the preprocessing stage. However, it is frequently thought of as little of and not broadly canvassed in writing. The point is to feature the significance of preprocessing strategies and show how they can improve framework accuracy. Specifically, some different preprocessing strategies are introduced, and the accuracy of each is contrasted and the others. The reason for this examination is to assess which methods are effective—likewise, present why the exactness improves, by methods for a detailed investigation of every strategy.

3. Materials and Methods

In figure 1, the complete workflow of our research is shown. First of all we imported all necessary libraries related to our research work in the Python platform. Later on, we will discuss those libraries one by one. After importing all mandatory libraries, the next step we performed is loading the dataset. In our research we worked on two different data sets. Both of the datasets are of social media. i.e. Tweeter. The data set contains views of people. It’s the time of technology so to get meaningful data from social media like reviews about the product, has become a challenging task [17]. Twitter contains a lot of data and mostly the data is noisy, contains urls and a lot of data that is of no use [18]. Still data is noisy and contains a lot of meaningless data, which consumes the memory and will slow down the processor speed. When this data set is loaded, the next step is to clean the data and make the dataset more precise. For this purpose, we performed various preprocessing steps. Those preprocessing steps are blank rows removal, stemming, punctuation marks removal, changing the text to lower case etc. After implementation of these steps the data is cleaned. Now corpus is generated for more accurate results. Then different machine learning algorithms are applied to this corpus. As a result of preprocessing, machine learning algorithms generated much better results as compared to the results which were generated without preprocessing.

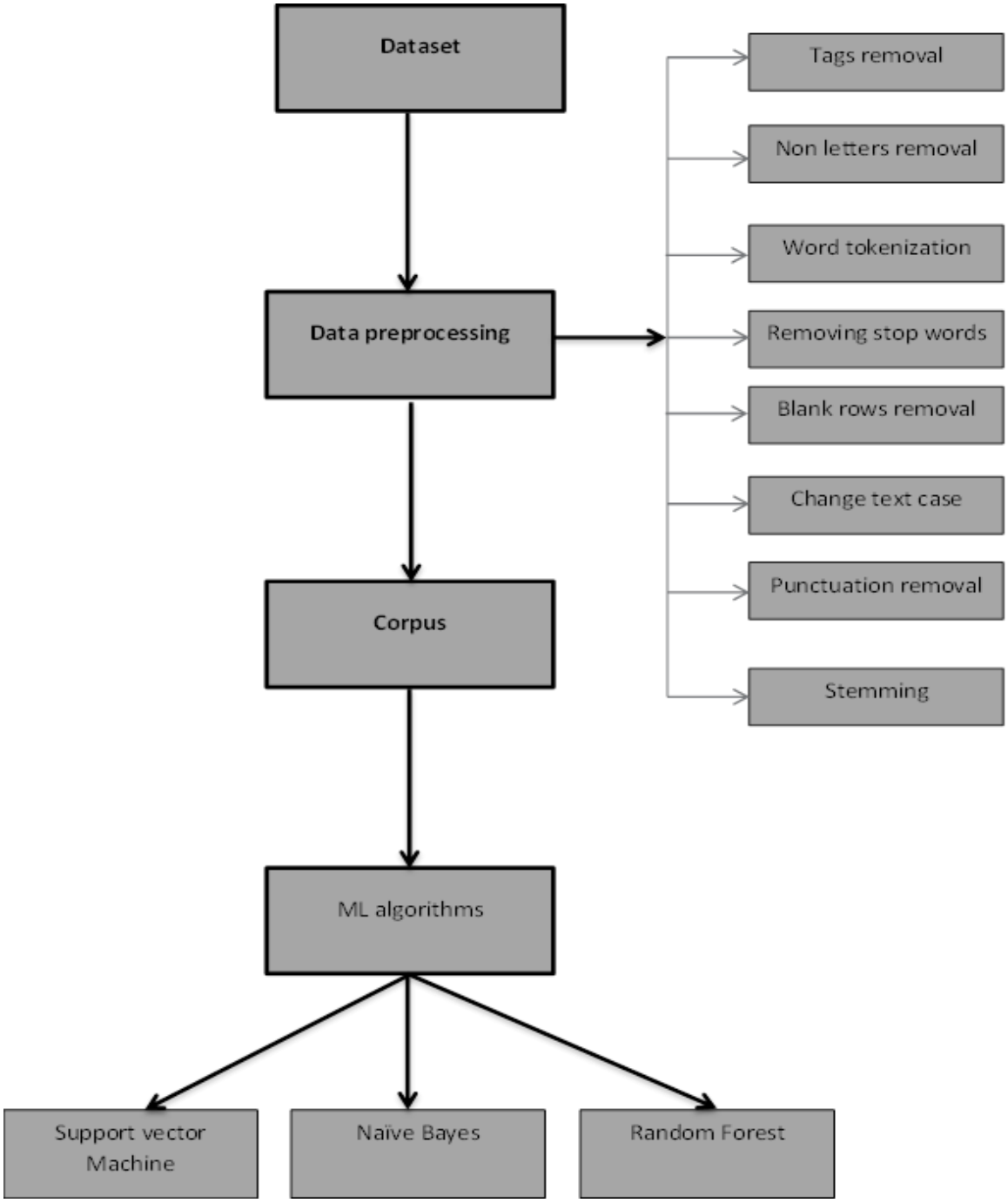


Figure 1: Workflow of proposed method

In this research we used the python platform with machine learning to get the desired results. The main focus of our work is NLTK (Natural Language Tool Kit). NLTK is a main stage for building Python projects to work with human language. Natural Language Tool Kit plays an important role in text preprocessing. In this research we performed work on two datasets taken from social media. Both of the data sets are different from each other. This research is based on [19]. In this research we find out the effect of different preprocessing steps on the

accuracy of machine learning algorithms. For this we chose Python language and later on applied different machine learning classifiers. In our research we used NLTK and BeautifulSoup mainly for preprocessing

After performing preprocessing, we applied machine learning algorithms. Machine learning algorithms include Support Vector Machine, Naïve Bayes and Random Forest. After applying all of these algorithms as classifiers, we got different results. These results told the accuracy of machine learning algorithms. On the basis of these outcomes, we can judge that to how much extent, a machine learning classifier can generate accurate results. Such results help out for prediction purposes.

3.1 *Framework*

We took two different datasets from Twitter and applied different preprocessing steps to make the data worth further processing. After preprocessing we applied different machine learning classifiers to get the accuracy. In our research we noticed that the preprocessing steps influenced the accuracy of machine learning algorithms. We applied three machine learning algorithms that are Support Vector Machine, Naïve Bayes and Random Forest.

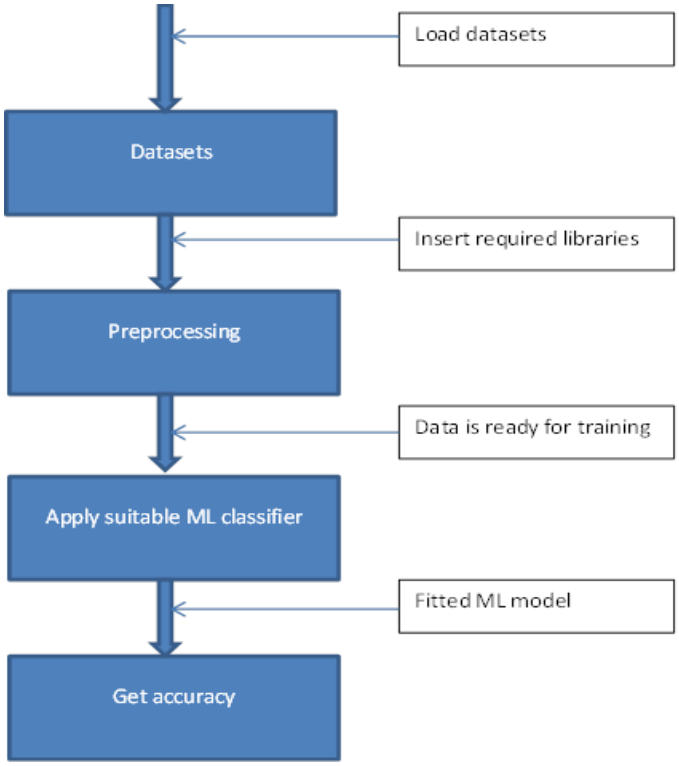
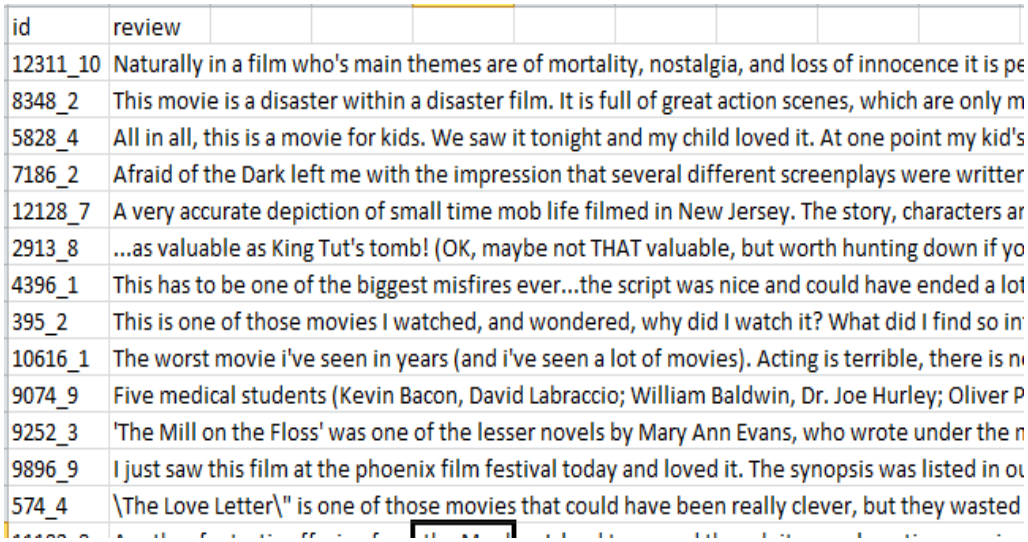


Figure 2: Framework of proposed work

3.2 *Datasets*

The collection of data is called a dataset. Dataset can be in numerous shapes like in tabular form, dataset also vary in size and styles. The data set in any research is of great importance. On the behalf of the data set, the final results are generated. One most important thing regarding data sets is that every dataset generates different and

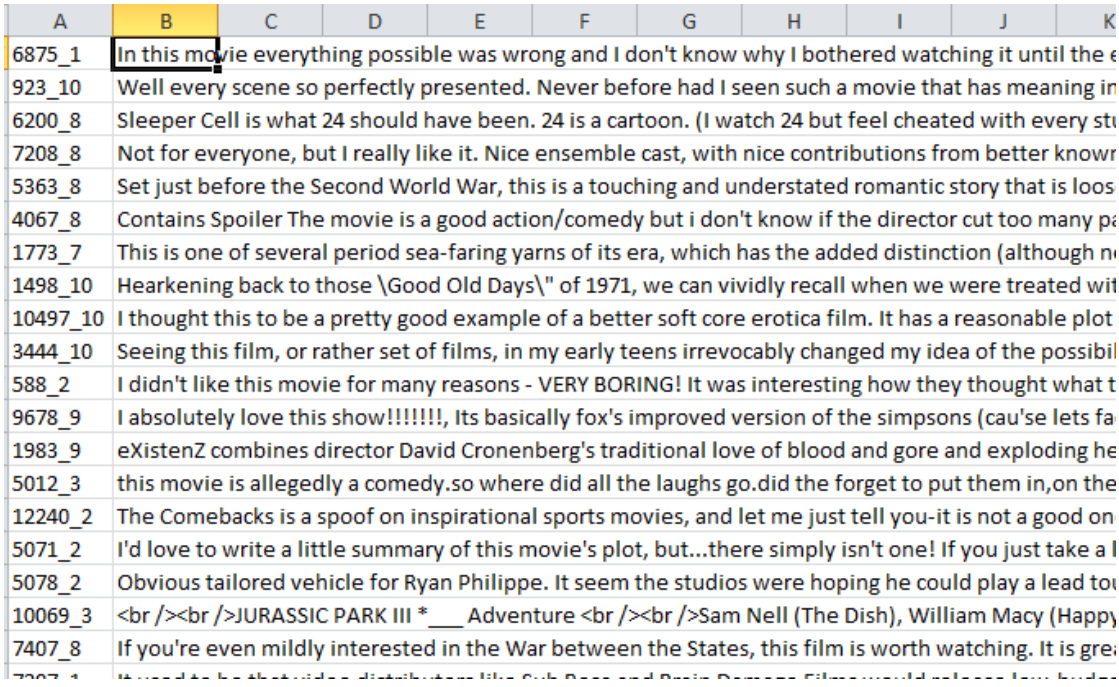
unique results. For the accuracy assessment of machine learning algorithms, dataset plays a vital role in generating results. The dataset is shown in figure 3.



id	review
12311_10	Naturally in a film who's main themes are of mortality, nostalgia, and loss of innocence it is pe
8348_2	This movie is a disaster within a disaster film. It is full of great action scenes, which are only m
5828_4	All in all, this is a movie for kids. We saw it tonight and my child loved it. At one point my kid's
7186_2	Afraid of the Dark left me with the impression that several different screenplays were writter
12128_7	A very accurate depiction of small time mob life filmed in New Jersey. The story, characters ar
2913_8	...as valuable as King Tut's tomb! (OK, maybe not THAT valuable, but worth hunting down if yo
4396_1	This has to be one of the biggest misfires ever...the script was nice and could have ended a lot
395_2	This is one of those movies I watched, and wondered, why did I watch it? What did I find so in
10616_1	The worst movie i've seen in years (and i've seen a lot of movies). Acting is terrible, there is n
9074_9	Five medical students (Kevin Bacon, David Labraccio; William Baldwin, Dr. Joe Hurley; Oliver P
9252_3	'The Mill on the Floss' was one of the lesser novels by Mary Ann Evans, who wrote under the n
9896_9	I just saw this film at the phoenix film festival today and loved it. The synopsis was listed in ou
574_4	\The Love Letter\" is one of those movies that could have been really clever, but they wasted

Figure 3: Dataset 1

In figure 3, a sample of our first dataset is shown. We also have the dataset of different categories. Here is given the sample of the other data set in the following figure 4.



A	B	C	D	E	F	G	H	I	J	K
6875_1	In this movie everything possible was wrong and I don't know why I bothered watching it until the e									
923_10	Well every scene so perfectly presented. Never before had I seen such a movie that has meaning in									
6200_8	Sleeper Cell is what 24 should have been. 24 is a cartoon. (I watch 24 but feel cheated with every stu									
7208_8	Not for everyone, but I really like it. Nice ensemble cast, with nice contributions from better knowr									
5363_8	Set just before the Second World War, this is a touching and understated romantic story that is loos									
4067_8	Contains Spoiler The movie is a good action/comedy but i don't know if the director cut too many pa									
1773_7	This is one of several period sea-faring yarns of its era, which has the added distinction (although n									
1498_10	Hearkening back to those \Good Old Days\" of 1971, we can vividly recall when we were treated wit									
10497_10	I thought this to be a pretty good example of a better soft core erotica film. It has a reasonable plot									
3444_10	Seeing this film, or rather set of films, in my early teens irrevocably changed my idea of the possibil									
588_2	I didn't like this movie for many reasons - VERY BORING! It was interesting how they thought what t									
9678_9	I absolutely love this show!!!!!!!, Its basically fox's improved version of the simpsons (cau'se lets fa									
1983_9	eXistenZ combines director David Cronenberg's traditional love of blood and gore and exploding he									
5012_3	this movie is allegedly a comedy.so where did all the laughs go.did the forget to put them in,on the									
12240_2	The Comebacks is a spoof on inspirational sports movies, and let me just tell you-it is not a good on									
5071_2	I'd love to write a little summary of this movie's plot, but...there simply isn't one! If you just take a l									
5078_2	Obvious tailored vehicle for Ryan Philippe. It seem the studios were hoping he could play a lead tou									
10069_3	 JURASSIC PARK III *___ Adventure Sam Nell (The Dish), William Macy (Happy									
7407_8	If you're even mildly interested in the War between the States, this film is worth watching. It is grea									

Figure 4: Dataset 2

3.3 *Text Preprocessing*

Whenever we have textual data, we have to apply a few pre-processing steps to the data to change words into mathematical highlights that work with machine learning calculations. The pre- processing ventures for an issue

rely essentially upon the space and the difficulty itself, henceforth, we don't have to apply all means to each issue. Preprocessing of text plays a vital role in producing the accuracy of machine learning algorithms [2]. We checked the accuracy of machine learning algorithms before and after applying different preprocessing techniques [25]. There are numerous preprocessing techniques. Every preprocessing step affects differently and gives different results.

In our research, we applied the following preprocessing steps:

- **Removal of emoticons**

  In this research, we are removing emoticons from our training dataset. Python supports fully removing emoticons. For this purpose, we are using the "emot. "library. Emot is a python library whose function is to extract emoji and emoticons from a textual string. These are the emoticons that are taken from the training dataset which was taken from Twitter. With the help of this python library, we cleaned our data from emoticons. Hence the training dataset is free of emoticons and generated accurate results as per expected demand. i.e import emot.

- **Removing non letters**

  Another preprocessing step is to remove all the digits or numbers in the data set. Our work in this research is purely based on textual data so any other kind of data like numbers or digits is completely removed from the dataset. The training dataset does not contain any digit in it. The presence of numbers in the data set affects the accuracy of results. Python provides numerous ways to remove numbers.

- **Word tokenization**

  In natural language processing, when we need each word for further analysis and process, word tokenization is applied. Word tokenization is referred to as the process in which a huge text is divided into small words. In this, each word is analyzed individually. For this purpose, we used NLTK (Natural Language Tool Kit) and divided a paragraph into words. Further, we installed "Punkt" for this function. The purpose of "Punkt" is to divide the whole paragraph into sentences.

- **Removing tags and markup**

  In the web world, when we want to move from one page to another, tags are used. These are the extra information in the text which contributes nothing while implementing the machine learning model. So these are extra data sources that's why we removed tags and markup from our data set. For this purpose, we imported the BeautifulSoup library and applied the method to remove tags and markup.

- **Removal of stop words**

  In our research stopwords are only useless data and add no contribution to the results. Stop words consume valuable space in the database. They also consume processing power and time. Their presence in the training dataset increases the size of the dataset and consumes more memory without any benefit. These words also affect the accuracy of machine learning algorithms. They are not giving any information about the results or any other thing. They only confuse the machine learning algorithms applied to the dataset.

- **Change the text to lower case**

  In this research, we changed the whole text of the training data set into lower case. The reason behind this is that Python interprets lower and upper words differently. For example, if there are two

words UNIVERSE and universe, python will understand that these two words are of different categories. So it's mandatory to convert all words into the same case.

- **Removal of punctuation**

  As we discussed above, anything except text creates problems while performing preprocessing. That's the reason we removed numbers from the data set and the training dataset is free of numbers. Same as numbers, punctuation marks also affect the accuracy of machine learning models. So we removed all punctuation marks from it. The training dataset is also free of punctuation marks.

- **Stemming**

  Stemming is a sort of normalization. It is used in the field of Natural Language Processing to prepare the text for further processing. Stemming refers to the derived words or sentences into their original form. It can also be referred to as some sort of normalization of text. Following modules are added for performing stemming: We first import the NLTK (Natural Language Tool Kit) library. From this library, we import "Porterstemmer ". In this research, we applied stemming techniques as a preprocessing step to test its effect on the accuracy of machine learning algorithms. There is a significant difference in the accuracy of machine learning algorithms with and without stemming implementation.

After implementation of preprocessing, corpus is generated.

3.4 *Machine Learning Algorithm*

Machine Learning is a subset of Artificial intelligence that is purely concerned with learning from its experience. Usually, we generate code to make computers learnable but in the case of machine learning, no coding is required in a sense to train the computer. Data is used as an input and from this input; the computer learns and makes a prediction. So we can say that in machine learning the decisions are data-driven

In our research we applied Naïve Bayes, Support Vector Machine and Random Forest.

Naïve bayes: Naïve Bayes is a classification technique that is purely based on the Bayes theorem. Bayes theorem tells how to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis, and the observed data itself. Implementation of naïve Bayes classifiers is easy and is applied to large datasets. Here is the mathematical description of naïve Bayes classifier:

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \qquad (1)$$

Where:

$$p(c|X) = p(x_1|c) \times p(x_2|c) \times \ldots \times p(x_n|c) \times p(c) \qquad (2)$$

Here P(c|x) is known as Posterior probability, P(c|x) as Likelihood, P(c) as Class prior probability, P(x) as Predictor prior probability.

Naïve Bayes classifiers work efficiently with high dimensions rather than low dimensions. Because the Naïve Bayes classifier is mostly used only for the calculation of probability, its implementation is comparatively easy. When we apply the Naïve Bayes classifier to independent assumptions, then the accuracy of this classifier is very high [20].

Support Vector Machine: Support vector machine is a machine learning algorithm which is useful in solving both the classification and regression problem [21]. In our research, the main aim of the Support Vector machine is to discuss classification problems.

Mathematical description of Support Vector Machine is as follows:

$$f(x) = sign\left(\sum_{i=1}^{i} a_i y_i k(x_{i,} x) + b\right) \quad (3)$$

Random forest: Random forest is a supervised learning technique which can be used for classification and regression. In our research, we used random forests for classification.As the name shows, in random forests there are many decision trees [23]. These decision trees are used to make a decision about any target attribute. So we can say that random forest is a kind of ensemble classifier. In ensemble classifiers the decision is taken on the bases of many supporting, not only depending upon a single entity.

3.5 *Python Libraries*

We also used many python libraries in our research work. The main python libraries used in our work are NLTK [24], Numpy, and Pandas.

4. **Results and Discussion**

In the current period, the online training framework produces an enormous measure of information on the web, generally in talk gatherings, coding assets, instructive locales, and papers. We utilized the unlabeled datasets and afterward utilized the datasets in the regulated learning model. In this work model, we utilized the accompanying technique. The data set which was initially unlabeled, we labeled it by supervised learning models. Our dataset consists of tweets containing reviews about a movie. We filtered out this data by different preprocessing steps. In this section we are discussing all the preprocessing performed and their results.

We took two different datasets from Twitter and applied different preprocessing steps to make the data worth further processing. After preprocessing we applied different machine learning classifiers to get the accuracy. In our research we noticed that the preprocessing steps influenced the accuracy of machine learning algorithms. We applied three machine learning algorithms that are Support Vector Machine, Naïve Bayes and Random Forest.

4.1 *Effect of Preprocessing on Naïve Bayes Accuracy*

In our research we applied Naïve Bayes algorithm on two different datasets taken from twitter. Both data sets gave different accuracy. Let us first discuss the accuracy of Naïve Bayes on the first data set. First of all we calculated the accuracy of Naïve Bayes algorithm before applying preprocessing steps. The accuracy was 79.345. Then we calculated the accuracy after performing preprocessing. The accuracy is 87.26. On the basis of this result we can say that accuracy of Naive Bayes classifiers increased significantly. Figure 5 shows the accuracy of the Naïve Bayes classifier.

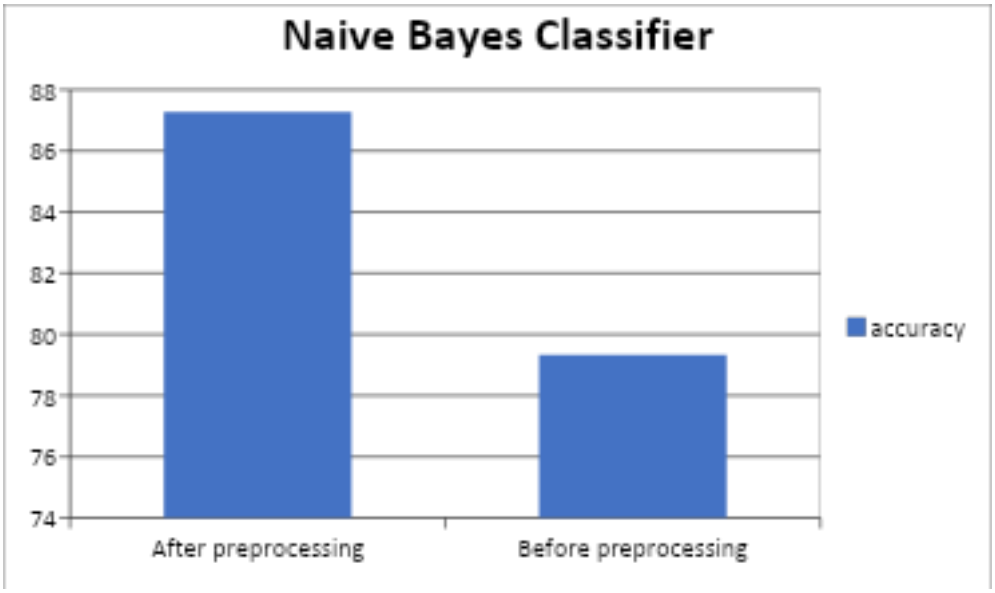


Figure 5: Effect of Preprocessing on Naïve Bayes

4.2 *Effect of Preprocessing on SVM Accuracy*

Now we will discuss the effect of preprocessing on the accuracy of the support vector machine. First of all we implemented a support vector machine on the Tweeter dataset without performing a single preprocessing step, then the accuracy we calculated is 76.52. Then we cleaned the data by applying different preprocessing techniques on the same dataset. The calculated accuracy after performing preprocessing is 77.62. Here figure 6 shows the result.

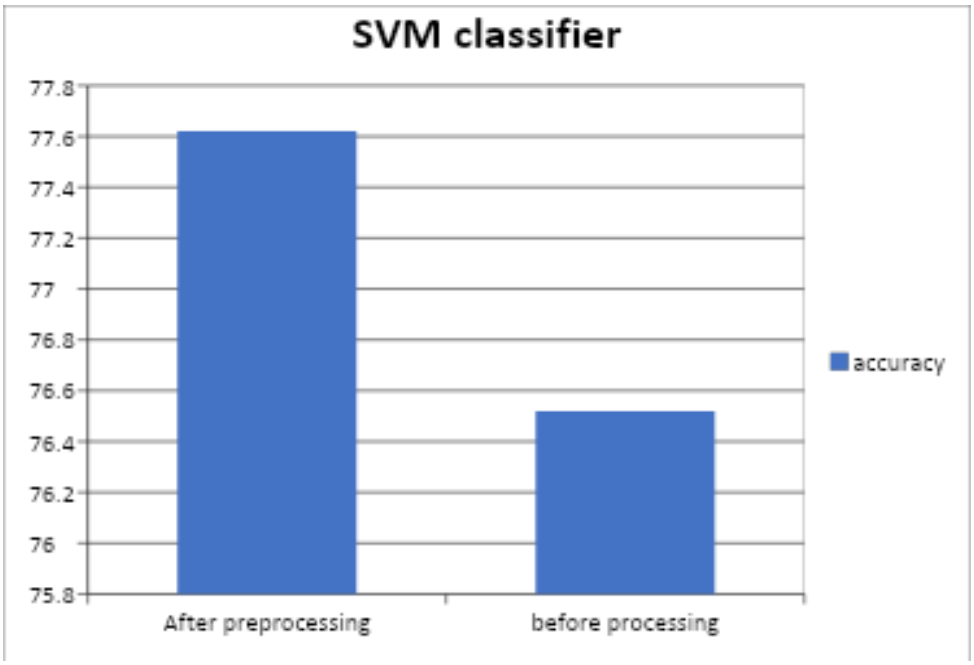


Figure 6:Effect of Preprocessing on SVM

We are going to discuss the results of each of the preprocessing steps individually.

4.3 *Evaluate the effect of removal of punctuation*

In this portion we are discussing the effect of punctuation removal on the accuracy of machine learning algorithms. We calculated the accuracy of SVM and Naïve Bayes before and after preprocessing. The accuracy of SVM before and after preprocessing is 77.389 and 75.99 respectively. This shows that by removing punctuation, the accuracy of SVM is decreased. Whereas, the accuracy of Naïve Bayes is increased from 60.839 to 61.072. Graphical representation of the results is shown in graph.

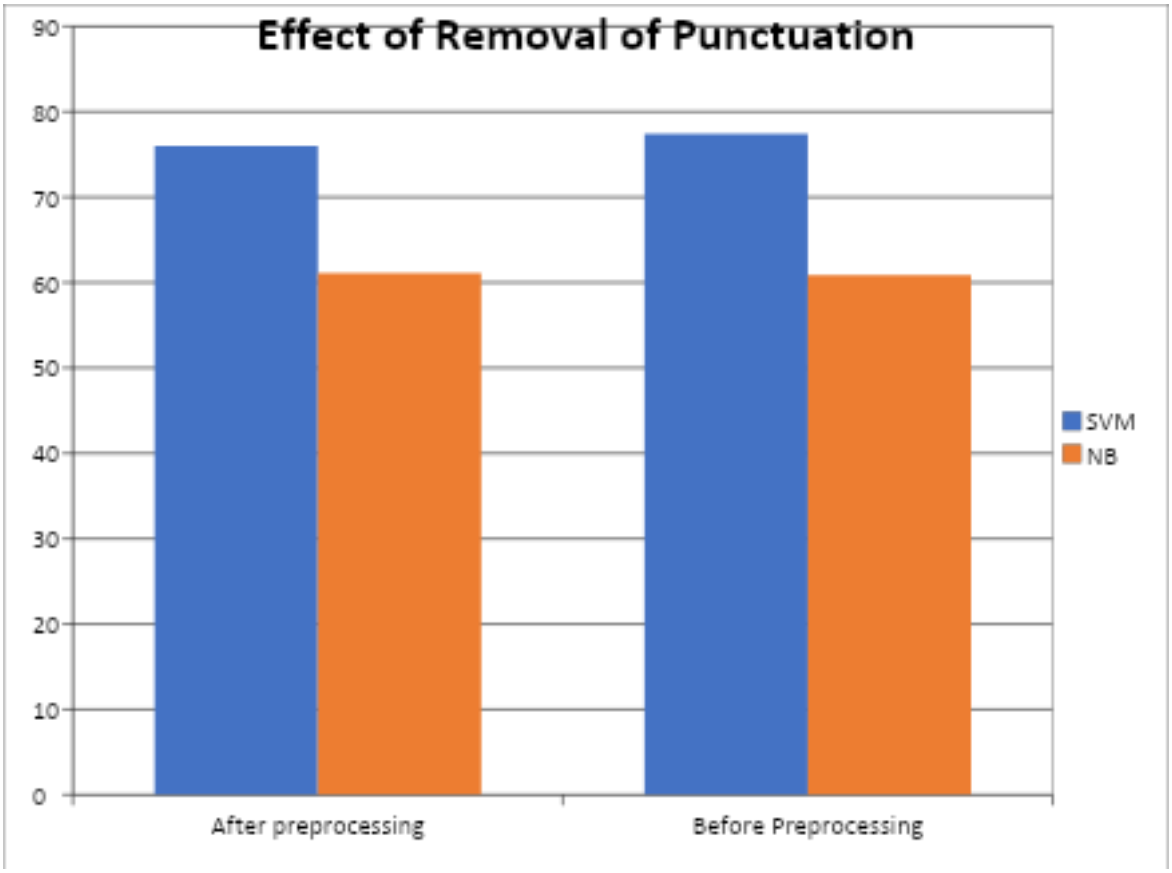


Figure 7: Effect of removal of punctuation

4.4 *Evaluate the effect of removal of Stop-Words*

Let us discuss the effect of removal of stop words on the accuracy of Naïve Bayes and SVM. Before applying the different preprocessing steps the accuracy of Naïve Bayes classifier was 60.839 and that of SVM was 77.389. After that we removed the stop words from the dataset. The calculated accuracy of SVM decreased to 75.52 and that of Naïve Bayes increased to 62.70. Graphical representation of these results is shown in the figure 7.

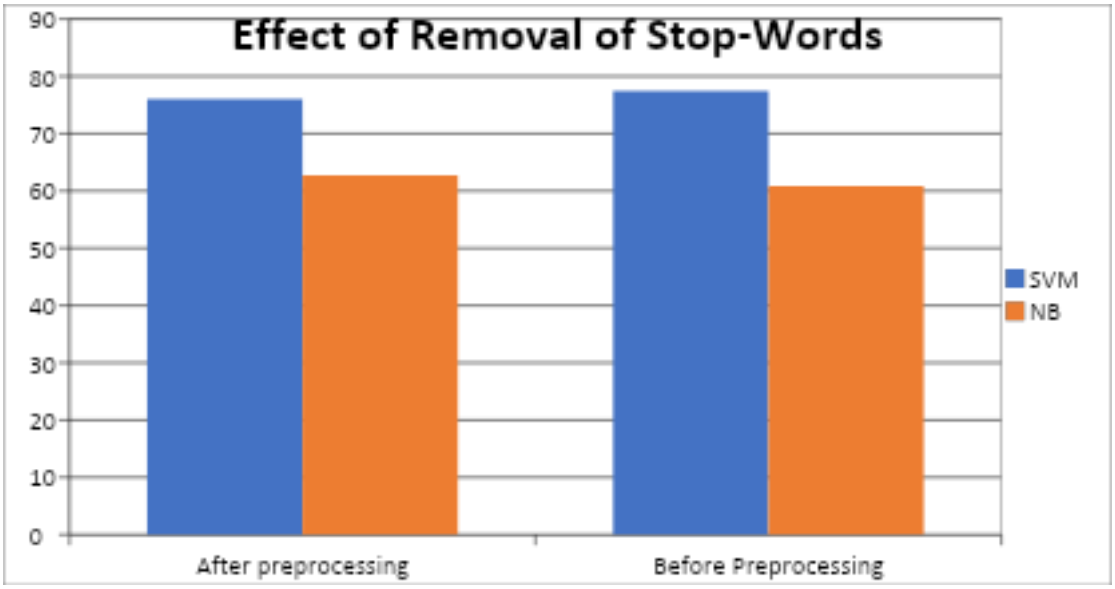


Figure 7: Effect of removal of Stop-words

4.5 *Effect of stemming*

Before implementation of stemming, we calculated the accuracy of machine learning algorithms. Before stemming, the accuracy of NB and SVM was 60.83 and 77.38 respectively shown in figure 8. After stemming, the calculated accuracy of SVM and Naïve Bayes is 77.62 and 60.66.

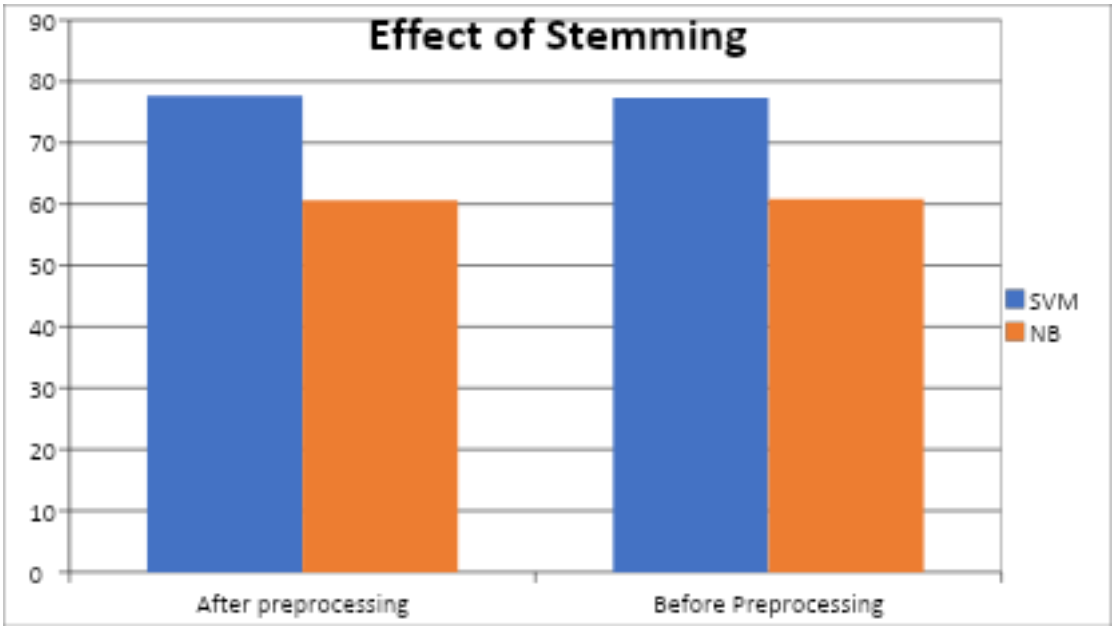


Figure 8: Effect of stemming

4.6 *Effect of removal of Blank rows*

Before removing blank rows from the dataset, we calculated the accuracy of machine learning algorithms. Before removing blank rows, the accuracy of NB and SVM was 60.83 and 77.38 respectively shown in figure 9. After removing the blank rows from the dataset their accuracy remains unaffected.

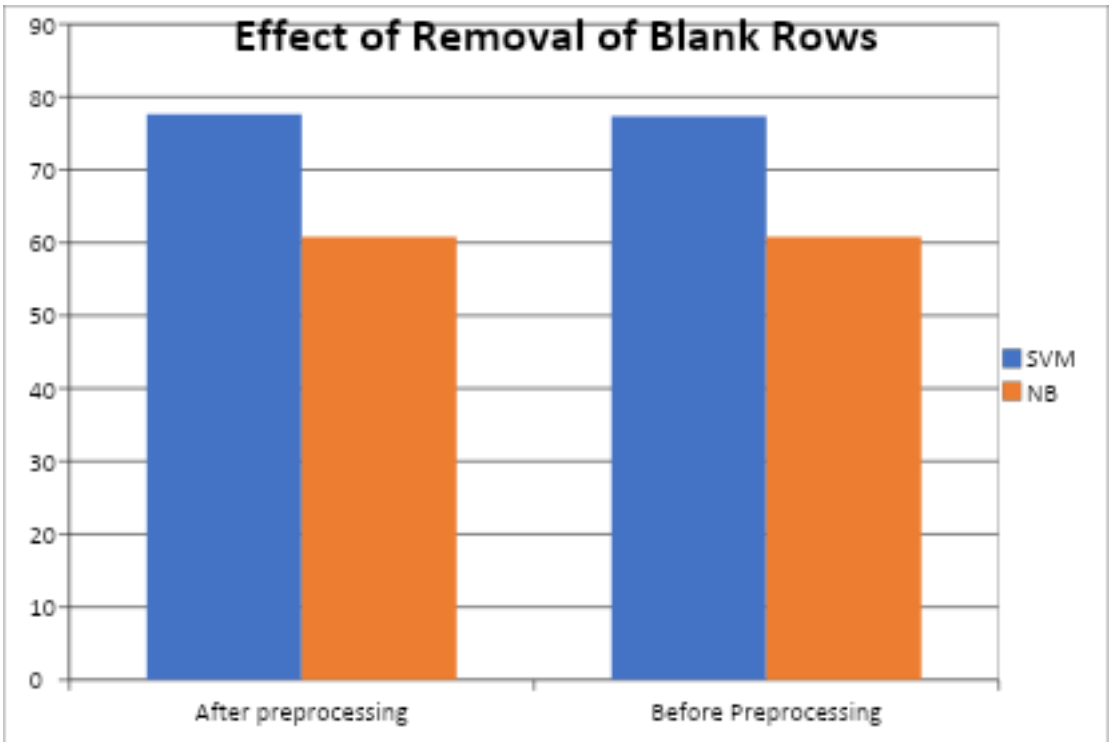


Figure 9: Effect of removal of blank-rows

5. Conclusion

In this research, preprocessing with a machine learning algorithm is the proposed methodology to meet the research objectives. In literary information science undertakings, any crude content should cautiously preprocess before the calculation can process it. We found the influence of preprocessing on the accuracy of machine learning algorithms. Furthermore we find out Fitting an appropriate model is necessary to get accuracy. Problems in fitting an appropriate model are under-fitting and over-fitting. An under-fit model will be less flexible and cannot account for the data. Over-fitting is a modeling error that occurs when a function is too closely fit to a limited set of data points. To solve the above mentioned problems, features and appropriate algorithms are selected.

An overall system for moving toward the textual information tasks comprises five stages: a collection of data, preprocessing of text, exploration and perception of the text, manufacturing the necessary model, and the last step is assessing the model. This system works in a loop format by nature. The accuracy of machine learning algorithms changes from data set to data set. Also accuracy changes with change of preprocessing techniques. At the end of the thesis, we computed accuracy results. All the accuracy results are different from each because accuracy changes from data set to data set. Another thing we noticed is that every single prepping technique influences classifiers in different ways.

References

[1] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," Comput. Math. Organ. Theory, vol. 25, no. 3, pp. 319–335, 2019, doi: 10.1007/s10588-018-9266-8.

[2] S. T. Ahmed, R. S. Al-Hamdani, and M. S. Croock, "EDM preprocessing and hybrid feature selection for improving classification accuracy," J. Theor. Appl. Inf. Technol., vol. 97, no. 1, pp. 279–289, 2019.

[3] P. Misra and A. S. Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," SSRN Electron. J., no. Ml, 2019, doi: 10.2139/ssrn.3349586.

[4] F. Z. Ruskanda, "Study on the Effect of Preprocessing Methods for Spam Email Detection," Indones. J. Comput., vol. 4, no. 1, p. 109, 2019, doi: 10.21108/indojc.2019.4.1.284.

[5] T. Iliou et al., "ILIOU machine learning preprocessing method for depression type prediction," Evol. Syst., vol. 10, no. 1, pp. 29–39, 2019, doi: 10.1007/s12530-017-9205-9.

[6] T. Dutoit, C. Martín-Vide, and G. Pironkov, "Preface," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11171 LNAI, pp. v–vi, 2018, doi: 10.1007/978-3-030-00810-9.

[7] Y. Singh and N. K. Garg, "Preprocessing Farmer Query Data Using Classic Method and Building Classifier Model," vol. 3, no. 3, pp. 1195–1199, 2018.

[8] A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," Int. J. Comput. Sci. Inf. Secur., vol. 16, no. 6, pp. 22–32, 2018.

[9] G. Orellana, B. Arias, M. Orellana, V. Saquicela, F. Baculima, and N. Piedra, "A study on the impact of pre-processing techniques in Spanish and english text classification over short and large text documents," Proc. - 3rd Int. Conf. Inf. Syst. Comput. Sci. INCISCOS 2018, vol. 2018-Decem, pp. 277–283, 2018, doi: 10.1109/INCISCOS.2018.00047.

[10] R. Hafeez, S. Khan, I. A. Khan, and M. A. Abbas, "Does preprocessing really impact automatically generated taxonomy," Proc. - 2017 13th Int. Conf. Emerg. Technol. ICET2017, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/ICET.2017.8281710.

[11] F. Mohammad, "Is preprocessing of text really worth your time for toxic comment classification?," 2018 World Congr. Comput. Sci. Comput. Eng. Appl. Comput. CSCE 2018 - Proc. 2018 Int. Conf. Artif. Intell. ICAI 2018, pp. 447–453, 2018.

[12] M. J. Denny and A. Spirling, "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It," Polit. Anal., vol. 26, no. 2, pp. 168–189, 2018, doi: 10.1017/pan.2017.44.

[13] J. Camacho-Collados and M. T. Pilehvar, “On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis,” arXiv, 2017, doi: 10.18653/v1/w18-5406.

[14] M. Kantepe and M. C. Gañiz, “Preprocessing framework for Twitter bot detection,” 2nd Int. Conf. Comput. Sci. Eng. UBMK 2017, pp. 630–634, 2017, doi: 10.1109/UBMK.2017.8093483.

[15] P. Chandrasekar and K. Qian, “The Impact of Data Preprocessing on the Performance of a Naïve Bayes Classifier,” Proc. - Int. Comput. Softw. Appl. Conf., vol. 2, pp. 618–619, 2016, doi: 10.1109/COMPSAC.2016.205.

[16] R. Asgarnezhad, M. Shekofteh, and F. Z. Boroujeni, “Improving diagnosis of diabetes mellitus using combination of preprocessing techniques,” J. Theor. Appl. Inf. Technol., vol. 95, no. 13, pp. 2889–2895, 2017.

[17] D. Munková, M. Munk, and M. Vozár, “Data pre-processing evaluation for text mining: Transaction/sequence model,” Procedia Comput. Sci., vol. 18, pp. 1198–1207, 2013, doi: 10.1016/j.procs.2013.05.286.

[18] E. Haddi, X. Liu, and Y. Shi, “The role of text pre-processing in sentiment analysis,” Procedia Comput. Sci., vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[19] M. Khanum, T. Mahboob, W. Imtiaz, H. Abdul Ghafoor, and R. Sehar, “A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance,” Int. J. Comput. Appl., vol. 119, no. 13, pp. 34–39, 2015, doi: 10.5120/21131-4058.

[20] Y. Tian, Y. Shi, and X. Liu, “Recent advances on support vector machines research,” Technol. Econ. Dev. Econ., vol. 18, no. 1, pp. 5–33, 2012, doi: 10.3846/20294913.2012.661205.

[21] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues, 9(5), 272–278.

[22] A. Kadhim, I. (2018). An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security, 16(6), 22–32.

[23] P. Kaviani, & S. Dhotre, (2017). International Journal of Advance Engineering and Research Short Survey on Naive Bayes Algorithm. International Journal of Advance Engineering and Research Development, 4(11), 607–611.

[24] E. Loper, & S. Bird, (2002). NLTK: The Natural Language Toolkit. https://doi.org/10.3115/1225403.1225421

[25] Omar, H., Dahab, M., & Kamal, M. (2016). Stemmer Impact on Quranic Mobile Information Retrieval Performance. International Journal of Advanced Computer Science and Applications, 7(12), 135–139. https://doi.org/10.14569/ijacsa.2016.071218