Original Article

WEIGHT BASED PRE-PROCESSING ALGORITHM (WBPA) FOR DATA PRE-PROCESSING IN SENTIMENT ANALYSIS

K. Brindha[1], Dr. E. Ramadevi[2]

[1]Research Scholar, NGM College, Tamilnadu, India, brindharajmohan004@ gmail.com

[2]Associate Professor, Dept of Computer Science, NGM College, Tamilnadu, India, ramasai1970@gmail.com

DOI: 10.47750/pnr.2022.13.S10.134

Abstract

In the big data period, data is made ceaselessly or nearer to constant. Social media, like Twitter, makes a gigantic proportion of such data. Regardless, social media data are every now and again unstructured and testing to make due. Thusly, this paper proposes a convincing text data pre-processing technique Weight Based Pre-processing Algorithm (WBPA) to deal with yahoo finance data. Foster an algorithm that weights the feeling score similarly as weight of hashtag and cleaned text and foster an algorithm to weight the scores of the hashtag and cleaned text to secure the opinion. The results show that stemming technique performed best with respect to computational speed. Besides, the accuracy of the algorithm was attempted against actually organized sentiments and sentiments conveyed before text data pre-processing. To the extent that model execution, the proposed algorithm performed better with the higher accuracy perhaps, due to the unstructured idea of yahoo finance data.

KEYWORDS: Social media data, Twitter, Machine Learning, text data Pre-processing, sentiment analysis, yahoo finance data.

1. INTRODUCTION

Internet data grow rapidly, given the inclination of occupants to share their viewpoints. Through the extension of social media, people's opinion contraptions have been refreshed, and fields, for instance, opinion mining and feeling investigation have procured developing requests. Online surveys can't be ignored inferable from the potential effects that client info could have on organizations [9]. A basic number of exploration practitioners are at present developing plans that can accumulate information from such contribution to help advancing figuring out, drive public feeling, and further develop client reliability. Subsequently, opinion examination was executed and applied in a new report locales and organizations. Twitter has been one of the most by and large used microblogging administrations and a fascinating conversation for more than 500 million messages every day from around 1.3 billion people. A message on Twitter (like a post on Facebook), by a progression of characters bound to 280-character limit, is posted openly or by spread out allies inferable from the record's security, which isn't exactly equivalent to other social media websites [12].

Twitter is a microblogging and progressing correspondence organization used by a great many people and associations to share and track down information. Clients can post their messages called 'tweets' which truly rely upon 140 characters [3]. Clients by and large present their tweets on offer their viewpoints or sentiments about items, occasions or people of note which can be positive, negative or neutral. For feeling investigation applications, twitter is the key wellspring of information and for examination of this data different AI techniques were used lately. In feeling examination, the demonstration of emojis acquires an excellent acclaim particularly in the midst of young people to communicate their difficult situation, fulfilment or shock towards an occasion, item or personality. Detecting and analyzing the uses of emojis acquired the assembly of experts in different areas of programming, clinical and social audit. Twitter broadcasts clients' short messages with respect to different pieces of life wherein a part of the tweets are significant for finding the information and opinion mining. Associations use these feedbacks for prediction and social issue supportive information.

Big data pre-processing is a vital endeavour for certain specialists, heads, associations and organizations to social event the data and exploring the tremendous proportion of explicit data or information. The different sorts of data or relative data are accumulated from different kinds of web sources and places. That data can be anticipated to measure in the term of boisterous data, missing data, wrong data or conflicting data inappropriately. Accepting the difference data give a few unacceptable results

wrong finishes in the data analysis, plan updates and bearing. The unique challenges are dealt with in blended variety data and unstructured data solicitation to be pre-handled into coordinated data or mentioned data representations. A movement of instruments, techniques and methodologies used to recognize and extricate information is described as sentimental analysis. The information extricated consolidates attitudes and opinion of the clients. The essential goal is to know whether the client has negative, positive and neutral opinion towards an item or something else. The amounts of papers on sentimental analysis have been extended definitely nowadays. It is one of the fastest developing assessment areas. It is seen that the general opinions, their motivation are political in nature.

Nonetheless, the natural language processing has been resolving various issues to congruity of sentimental analysis. The undertakings have been advanced from fundamental limit location to complex emojis, isolating negative feelings. Navigation is one of the critical elements "what others are thinking". For example while needing to project a ballot in races. By utilizing internet looking through the singular's opinion from personal network is conceivable. Besides, various people are making their opinions through internet, for example, from comments, blogs in social networking objections. Gigantic piece of our general public is related through social networks straightforwardly or by implication. Anyone can offer their viewpoints without any feeling of fear toward lamentable results. To construe the opinion of clients from extensive comments, casual language, for example, emojis, compression will befuddle the analysis. In this paper, AI algorithms were applied with optimization technique's to isolate features and opinions.

1.1 Sentiments

Sentiment analysis is a computational cycle which perceives and sorts an opinion in a piece of text that communicates the positive, negative, or neutral demeanor of a writer towards a particular item, occasion or personality. For example, consider consolidates tweets which are named positive, negative, or neutral. Some of the time sentiments are questionable; that either a tweet is positive or negative [2]. Such sentiments are called neutral sentiments. To avoid ambiguity in our assessment work eliminated the neutral sentiments. Utilized a similar procedure to kill neutral sentiments, in other words, a tweet should be considered neutral expecting it appears as a title on the main page of a newspaper or a declaration in Wikipedia.

1.2 Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to decide if data is positive, negative or neutral. Sentiment analysis included distinguishing a given text of content by first pre-processing it to detecting stop words and images, and so on and afterward checking the subjectivity contents [1].

Sentiment classifies the substance into positive or negative or potentially neutral. SA utilizes information in the term of context-subordinate, for instance, a few single words gave various importance in the given word. Sentiment analysis is a fundamental undertaking to distinguish the sentiment polarities in the text applied broadly in online business framework, blogs, and social media. Its fundamental undertaking bunches the archive into different polarities. In view of automatic prediction, the brokers can pursue choice more straightforward, and furthermore plan the bearing to foster their business.

The customary methodology is typically supervised learning, supervised classifiers are utilized like Naive Bayes, SVM, logistic regression, gathering of casting a ballot classifiers, additionally researching on include choice for holding valuable features and ignoring redundant features to further develop the performing approach. This relies upon the size and nature of the pre-marked datasets which are scant and inaccessible for a specific application, they are dreary to gather, expensive and time-consuming to construct, rely upon space transformation and insufficiently handle inconspicuous data. Particularly, the Vietnamese preparation data are not plentiful and need so much preventing many suggestions in research group. This spurs us to foster a successful answer for text characterization overall and sentiment analysis specifically.

Sentiment analysis fills in as a central part of managing customers on web-based entrances and websites for the organizations. They do this all the time to characterize a remark as a query, complaint, suggestion, opinion, or just love for an item. This way they can without much of a stretch sort through the comments or questions and prioritize what they need to deal with first and even request them such that is more appealing. Organizations sometimes even attempt to erase content that has a negative sentiment joined to it.

1.3 Machine Learning

Machine learning is a discipline that empowers a PC to learn without expressly in the program. In machine learning algorithms can be gathered in view of anticipated information and result of the algorithm.

1. Supervised learning, make a capability which guides contribution to yield wanted. Like grouping (classification) a learning algorithm depends on an example set of information yield matches are beneficial in enormous enough amounts. This algorithm notices these models and afterward delivers a model equipped for planning the new contribution to the fitting result.

2. Unsupervised learning, modeling the arrangement of sources of info, like classification (clustering). This algorithm has the goal to read up and search for fascinating examples on a given info. Albeit not gave appropriate result unequivocally. One of unsupervised learning algorithms most regularly utilized is clustering or grouping.

2. LITERATURE SURVEY

2.1 Text Pre-processing in Sentiment Analysis

Emma Haddi et.al proposed The Role of Text Pre-processing in Sentiment Analysis. It is trying to understand the most recent patterns and synopses the state or general opinions about items because of the big variety and size of social media data, and this makes the need of mechanized and constant opinion extraction and mining. Mining on the web opinion is a type of sentiment analysis that is treated as a troublesome text classification task. In this paper, they investigate the job of text pre-processing in sentiment analysis, and report on trial results that exhibit that with suitable component determination and representation, sentiment analysis correctnesses utilizing support vector machines (SVM) in this space might be essentially moved along. That could profit from its outcomes, for example, news analytics, marketing, question answering, readers do. Hearing significant bits of knowledge from thoughts expressed on the internet particularly from social media blogs is imperative for some organizations and establishments, whether it is regarding item feedback, public mood, or financial backer's opinions.

2.2 Data Pre-Processing for Machine Learning Models using Python Libraries

Namrata Pandey et.al proposed Data Pre-Processing for Machine Learning Models using Python Libraries. Data pre-processing is the method involved with transforming the crude data into helpful dataset. Data pre-processing is one of the main period of any machine learning model on the grounds that the quality and productivity of any machine learning model straightforwardly relies on the data-set, in the event that they skirt this step and plan a model with data sets containing missing qualities then the model they have planned won't be that effective and will be conflicting model. This paper portrays the methodology for pre-processing the data in seven grouping of steps utilizing python strong libraries which are open source machine learning libraries that support both supervised and unsupervised learning like pandas is a general data control tool, scikit learn which gives different tools to show fitting, data pre-processing, model choice and numerous different utilities. These means incorporate managing missing worth, unmitigated qualities, bringing in data sets and so on. They have likewise taken care of the downright factors, partition of free factor and ward variable. Additionally, dataset is demonstrated disdain toward as training and testing datasets.

2.3 Pre-processing input data for machine learning by FCA

Jan Outrata et.al proposed pre-processing input data for machine learning by Formal concept analysis (FCA). Two pre-processing methods are presented. The first comprises in expanding the arrangement of qualities portraying objects in input data table by new traits and the subsequent one comprises in supplanting the properties by new characteristics. In the two methods the new properties are characterized by specific formal concepts figured from input data table. Chosen formal concepts are purported factor concepts got by Boolean factor analysis, as of late depicted by FCA. The ML strategy used to show the thoughts is choice tree acceptance. The exploratory assessment and correlation of execution of choice trees incited from unique and pre-handled input data is performed with standard choice tree enlistment algorithms ID3 and C4.5 on a few benchmark datasets. The two methods use Boolean factor analysis, as of late portrayed by FCA, in that the new properties are characterized as factors figured from input data. The number of factors is generally more modest than the number of attributes.

2.4 Data Pre-processing and Intelligent Data Analysis

A. Famili et.al proposed Data Pre-processing and Intelligent Data Analysis. The paper examines exhaustively two primary explanations behind performing data pre-processing: (I) problems with the data and (ii) preparation for data analysis. The paper goes on with subtleties of data pre-processing techniques accomplishing every one of the previously mentioned targets. A sum of 14 techniques is examined. Two instances of data reprocessing applications from two of the most data rich spaces are given toward the end. The applications are connected with semiconductor assembling and aviation spaces where a lot of data are accessible and they are genuinely reliable. Future headings and a few challenges are examined toward the end. Perhaps of the main issue in data pre-processing is how they have any idea about what valuable information exists in the crude data with the goal that they can ensure it is preserved. This might rely on our meaning of data pre-processing. Some might contend that data pre-processing isn't a totally "pre-" interaction of data analysis. It needs feedback from the principal data analysis process. All things considered, a definitive judgment of whether one has worked effectively of data pre-processing is to check whether the "valuable information" has been found in the later data analysis process.

2.5 Universal Language Model Fine-Tuning and SVM

Barakat AlBadani et.al proposed A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM. Twitter sentiment detectors (TSDs) give an improved answer for assess the nature of administration and item than other customary technologies. The classification accuracy and detection execution of TSDs, which are very dependent on the exhibition of the classification techniques, are utilized, and the nature of input features is given. In any case, the time required is a big issue for the current machine learning methods, which prompts a test for all endeavors that plan to change their organizations to be handled via computerized work processes. Deep learning techniques have been used in a few certifiable applications in various fields like sentiment analysis. Deep learning approaches utilize various algorithms to get information from crude data like texts or tweets and represent them in particular kinds of models. These models are utilized to deduce information about new datasets that poor person been modelled at this point. They present another viable strategy for sentiment analysis utilizing deep learning designs by consolidating the "universal language model fine-tuning" (ULMFiT) with support vector machine (SVM) to build the detection efficiency and accuracy. The strategy presents another deep learning approach for Twitter sentiment analysis to distinguish the attitudes of individuals toward specific items in light of their comments. The proposed model presents a successful deep learning design that consolidates the universal language model fine-tuning with a support vector machine. The broad outcomes on three genuine world datasets show that the proposed model increments detection efficiency and accuracy.

2.6 Parts of Speech Tagging

Saritha Shetty et.al proposed Text pre-processing and parts of speech tagging for Kannada language. Part-of-Speech (POS) tagging is one of the most important tasks in the field of natural language processing (NLP). POS tagging for a word depends not only on the word itself but also on its position, its surrounding words, and their POS tags. POS tagging can be an upstream task for other NLP tasks, further improving their performance. The technique of assigning various parts of speech for each word in a text file is referred as Parts of Speech Tagging. This paper propounds assigning of POS for Kannada Language words by Hidden Markov Model. This paper also focuses on Kannada Language detection and Text pre-processing. POS tagger has been developed using Python programming language. Tkinter is used as an interface. Data accumulation for training and testing of the system is done from wikipedia, Kannada e-papers. 18000 words are trained and they are tested with 1000 words. The contrast amidst project generated output and physically tagged data results in correctness of accuracy for POS tagging. The correctness of 95% is achieved from the experimental outcome of the proposed System.

3. PROPOSED METHODOLOGY

3.1 Data-sets

The Data sets are taken by real word datasets; this research scrapped the data from finance yahoo stocks and publicly available Twitter data. This incorporates the timestamp and tweet text for every tweet of a particular period. Since predictions are being made on daily basis, tweets are split by day using their timestamps.

3.2 Pre-processing of tweets

Raw tweets collected from kaggle website generally contain noisy data. Hence, clean the raw yahoo finance data before using them to train various classifiers. Have performed an extensive pre-processing to prepare the dataset and to shorten the size of the dataset and first perform some typical pre-processing steps as described below:

I) Replace two or more dots (.) with the same number of spaces.

II) Remove spaces and quotes (" and ') from the ends of tweets.

III) Replace two or more spaces with a single space.

IV) Remove the repetition characters more than two characters back to two.

V) Replace numbers and special characters with spaces.

In addition, clean the tweets using few more techniques:

Re-mapping Class Labels and Resizing Dataset

Firstly, change the sentiment classification from (0 negative, 4 positive) to (0 negative, 1 positive) and remove the column such as Username, Query, Date, and ID.

HTML Decoding

Performed HTML decoding to remove text fields such as 'amp', 'quot', etc, used Beautiful Soup python package to perform this decoding.

Dealing with @text

Have removed @text even though @text contains specific information and this generally does not contribute positively to build a sentiment analysis model.

Converting upper case to lower case

In care using case sensitive analysis, might take two occurrences of same words as different due to their sentence case it important for an effective analysis not to provide such misgivings to the model.

Removing URL Links

To classify and to analyze the sentiment of the tweets, URLs will not be considered usually. For example 'your chat is not working on your site: http://ecstacy.com as she is feeling excited". In this URL by considering the word 'not' it can be referred to a negative sentiment but the word excited refers to a positive sentiment. Hence it will be considered as neutral. In order to remove such kind of sentiments the URLs are removed.

Removal of hash symbols

They are unspaced phrases prefixed by the hash symbol (#). Users generally use them to indicate an important topic. They are replaced without the hashtags. For example, #cloud is replaced by cloud. These labels act as keywords and helpful in search of particular messages and posts. For example #guilty pleasures will produce tweets list which are related to #guilty pleasures. These are not needed in polarity detection. Hence considered as irrelevant and removed.

Handling Missing Values

After cleaning the dataset, found 2716 missing instances and have removed these instances since there is no text in them. So have in total 1597284 instances remaining in the dataset.

Removing retweets

Copying of another client tweets and posting it into one more record is considered as re tweeting. It is shortened as RT, to keep away from overt repetitiveness, re tweets are removed.

Removal of emoticons

Users usually use emojis to communicate their feelings; like smiley, sad, angry, and happy. For example consider the tweet ‘‘thoroughly enjoyed shovelling the driveway today! :)’’ It is hard to identify a user’s approach towards a product, personality or event to be positive or negative using emoticons in the post. In this research are removing emoticons from our training dataset.

Removal of stopwords

In English, stopwords are: the, is, at, which, on and so on. Filtered out stopwords from our dataset as they are conventionally high in frequency and are not giving any useful information in fact, they may puzzle a machine learning algorithm.

Stemming

Stemming is a technique used in information retrieval to combat the lexicon mismatch problem, in which a query’s words do not match with the words of a document. In English and numerous other western European languages, stemming is primarily a course of postfix deliberation. In microblogging, spelling mistakes and mismatches are normal, for example, @stellargirl I loooooooovvvvvveee my Kindle2. In training dataset there is no such word ‘‘love’’ with so many ‘‘os, vs and es’’ Used the stemming technique as a pre-processing step to improve the accuracy of MLAs.

To use different techniques, first, the data was passed to the common process, which consists of techniques mentioned in above. The sentiment obtained was then stored in the list. In the same time, the sentiments were converted to the numeric value by finding out which type of sentiment:

$$f(x) = \begin{cases} 0 & if x = negative \\ 1 & if x = neutral \\ 2 & if x = positive \end{cases}$$

Where x is the sentiment (positive, negative or neutral) in order to identify the performance of the technique was used.

3.3 Data Analysis

Data analysis was run on Lenovo G50 with the i7 processor running latest Windows 10 computer system. Appropriate libraries were installed to run the proposed WBPA programs. The data have been gathered by using a python library named GetOldTweets3. That data was pre-processed and then passed to the algorithm developed WBPA. The algorithm uses the weight of the hashtag and the weight of the cleaned tweet. The purpose of using the hashtag weight is because it is useful in a recommendation system, classification, categorization, and search. Hashtags (i) facilitate the search of topics based on social content with themes, (ii) provide support to the user to identify relevant topics, (iii) the recommendation system built using hashtag has also received much attention. All of which highlights the importance of using hashtags in our algorithm.

3.4 Proposed WBPA Algorithm

Develop a new algorithm to provide proportional weight between the hashtag and cleaned text combined to obtain sentiment output.

The weight is calculated using

$Tw=\alpha H+\beta T$

Where α weight of the hashtag and β is the weight of the tweets, H is the hashtag score, and β is the clean tweet score.

Algorithm for weighting the text

Step 1: Start the process
Step 2: While (tweet! = 0):
Step 3: Work out the tweet sentiment polarity
Step 4: If (hashtag != 0):
Step 5: Work out the hashtag sentiment polarity
Step 6: End if
Step 7: If (hashtag == 0):
Step 8: Set the weight of the tweets to100%
Step 10: End if
Step 11: Repeat the process
Step 12: Compute $Tw = \alpha H + \beta T$
Step 13: Allot the last polarity
Step 14: In light of polarity decide the sentiment
Step 15: Until there are no tweets left

Step 16: End the process.

Algorithm for weighting the text: While (tweet! = 0): Calculate the tweet sentiment polarity If (hashtag! = 0): Calculate the hashtag sentiment polarity End if (hashtag == 0): Set the weight of the tweets to 100%. End if Repeat: Compute $Tw=\alpha H + \beta T$ Assign the final polarity Based on polarity determine the sentiments Until there are no tweets left main subject of the text, has been shown to provide valuable information and has been used in sentimental analysis. For example, the proportional weight given to the hashtag in the algorithm is 40%, and the weight given to the cleaned text is 60%. If there is no hashtag in the tweet, the full weight (100%) is given to the cleaned text, and the sentiment is produced.

4. PERFORMANCE METRICS

4.1 Accuracy

Accuracy or Accuracy rate (or percent correct), is characterized as the quantity of right cases separated by the total number of cases. It is a measure of how regularly a sentiment rating was right. This is the most normally utilized measurement to assess how well the AI model is doing. Accuracy is the proportion of the quantity of right forecasts made against the complete number of expectations made. This is the estimation of the whole obvious anticipated worth against all the predicated esteem. Accuracy is measured in percentage and is registered.

$$\textbf{Accuracy} = \frac{\textbf{True positive} + \textbf{True Negative}}{\textbf{True positive} + \textbf{False negative} + \textbf{False positive} + \textbf{True Negative}}$$

4.2 Precision

Precision (additionally called positive predictive value) is the negligible part of recovered cases that are important or it is the level of chosen things that are right. It is the quantity of right positive outcomes partitioned by the quantity of positive outcomes anticipated by the classifier. A measure of the number of reports with feeling was evaluated as nostalgic. This could be viewed as how precisely the framework decides impartiality. For the most part, high recall scores are exceptionally troublesome in trial of wide topic, as the framework is needed to see ever-bigger arrangements of words and language. This is the estimation of genuine positive forecast against all positive expectation. Precision positive (p) and Precision negative (n) are precision proportion and are registered.

$$\textbf{Precision} = \frac{\textbf{True positive}}{\textbf{True positive} + \textbf{False positive}}$$

4.3 F-measure

It is additionally called F-Score or F-Measure; this is a mix of precision and recall. The score is in a scope of 0.0 - 1.0, where 1.0 would be awesome. F1 Score is the Harmonic Mean among precision and recall. It discloses to you how exact your classifier is, just as how powerful it is. High precision yet lower recall, gives you an amazingly exact, yet it at that point misses an enormous number of examples that are hard to order. The more prominent the F1 Score, the better is the exhibition of our model. F1 Score attempts to discover the harmony among precision and recall. The F1 Score is useful, as it gives us a solitary metric that rates a framework by both precision and recall.

$$\textbf{F} - \textbf{Measure} = \frac{\textbf{2} \times \textbf{precision} \times \textbf{Recall}}{\textbf{precision} + \textbf{Recall}}$$

4.4 Recall

"Recall" (additionally called Sensitivity, Hit Rate, and True Positive Rate) mentions to you what extent of information that really is positive was anticipated positive. In other words, the proportion of True Positive in the set of all actual positive data, Tracking down that semantic highlights produce better Recall and F score when grouping negative supposition, and better Precision with lower Recall and F score in positive conclusion arrangement. Recall is the best reasonable assessment strategies for text applications. They measure how exact and complete the order has been finished. Recall is called as the no of genuine positive isolated by indisputably the no of segments that are successfully having a spot with the positive class.

$$\textbf{Recall} = \frac{\textbf{True positive}}{\textbf{True positive} + \textbf{False negative}}$$

5. EXPERIMENT RESULTS

5.1 Accuracy

Table 2.Comparison table of Accuracy

Data sets	POS-Tagging	SVM	FCA	Proposed Pre-processing using WBPA
10	0.27	0.30	0.32	0.39
25	0.38	0.42	0.44	0.51
50	0.50	0.60	0.55	0.69

75	0.57	0.75	0.61	0.77
100	0.78	0.80	0.84	0.92

The Comparison table 2 of Accuracy Values explains the different values of existing algorithms (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA. While comparing the Existing algorithm (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA, provides the better results. The existing algorithm values start from 0.27 to 0.78, 0.30 to 0.80, 0.32 to 0.84 and proposed Pre-processing WBPA values starts from 0.39 to 0.92.

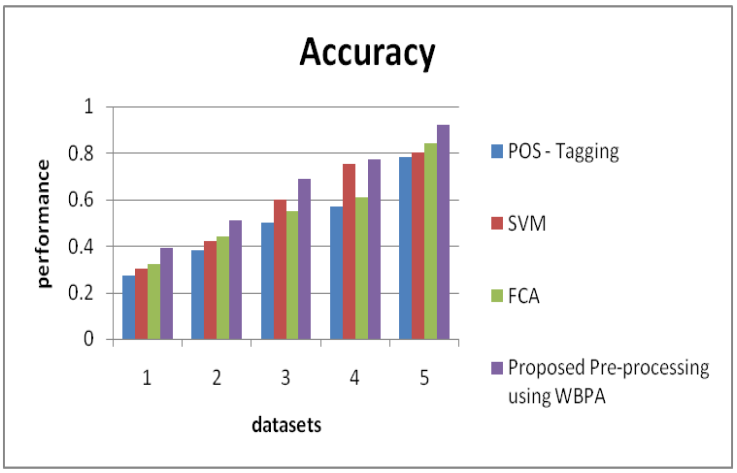

Figure 2. Comparison chart of Accuracy

In Figure 2 illustrates the comparison chart of Accuracy demonstrates the existing1, existing 2 (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA values. X axis denote the datasets and y axis denotes the performance values in accuracy. The existing algorithm values start from 0.27 to 0.78, 0.30 to 0.80, 0.32 to 0.84 and proposed Pre-processing using WBPA values starts from 0.39 to 0.92, which provides the great results.

5.2 Precision

Table 3.Comparison table of Precision

	POS-Tagging	SVM	FCA	Proposed Pre-processing using WBPA
Positive	0.30	0.36	0.32	0.39
Negative	0.41	0.46	0.44	0.51
Other	0.75	0.85	0.81	0.88

The Comparison table 3 of precision Values explains the different values of existing algorithms (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA. While comparing the Existing algorithm (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA, provides the better results. The existing algorithm values start from 0.30 to 0.75, 0.36 to 0.85, 0.32 to 0.81 and proposed Pre-processing using WBPA values starts from 0.39 to 0.88.

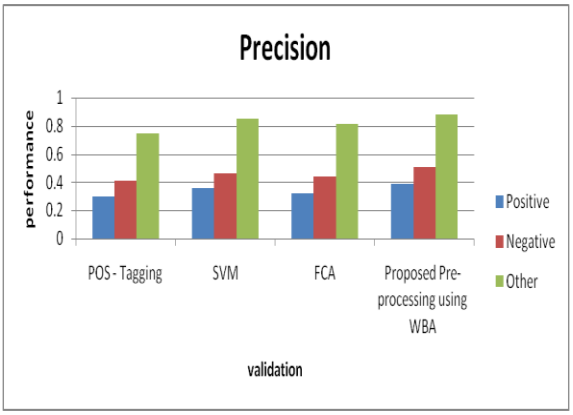

Figure 3. Comparison chart of Precision

In Figure 3 illustrates the comparison chart of precision demonstrates the existing1, existing 2 (POS-Tagging, SVM, FCA) and proposed Pre-processing using WBPA. X axis denote the validation and y axis denotes the performance values in precision. The existing algorithm values start from 0.30 to 0.75, 0.36 to 0.85, 0.32 to 0.81 and proposed Pre-processing using WBPA values starts from 0.39 to 0.88, which provides the great results.

5.3 Recall

Table 4.Comparison table of Recall

	POS-Tagging	SVM	FCA	Proposed Pre-processing using WBPA
Positive	0.32	0.60	0.36	0.47
Negative	0.41	0.65	0.62	0.53
Other	0.50	0.43	0.70	0.86

The Comparison table 4 of precision Values explains the different values of existing algorithms (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA. While comparing the Existing algorithm (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA, provides the better results. The existing algorithm values start from 0.32 to 0.50, 0.60 to 0.43, 0.36 to 0.70 and proposed Pre-processing using WBPA values starts from 0.47 to 0.86. Proposed algorithm provides the great results.

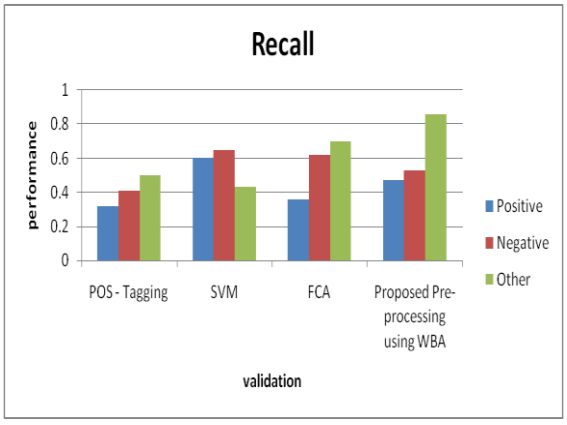

Figure 4.Comparison chart of Recall

In Figure 4 illustrates the comparison chart of Recall demonstrates the existing1, existing 2 (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA. X axis denote the validation and y axis denotes the performance values in Recall. The existing algorithm values start from 0.32 to 0.50, 0.60 to 0.43, 0.36 to 0.70 and proposed Pre-processing using WBPA values starts from 0.47 to 0.86. Proposed algorithm provides the great results.

5.4 F1- score

Table 5.Comparison table of F-Measure

	POS-Tagging	SVM	FCA	Proposed Pre-processing using WBPA
Positive	0.28	0.30	0.33	0.41
Negative	0.40	0.45	0.43	0.49
Other	0.56	0.60	0.74	0.81

The Comparison table 5 of precision Values explains the different values of existing algorithms (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA. While comparing the Existing algorithm (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA, provides the better results. The existing algorithm values start from 0.28 to 0.56, 0.30 to 0.60, 0.33 to 0.74 and proposed Pre-processing using WBPA sentence scoring values starts from 0.41 to 0.79. Proposed algorithm provides the great results.

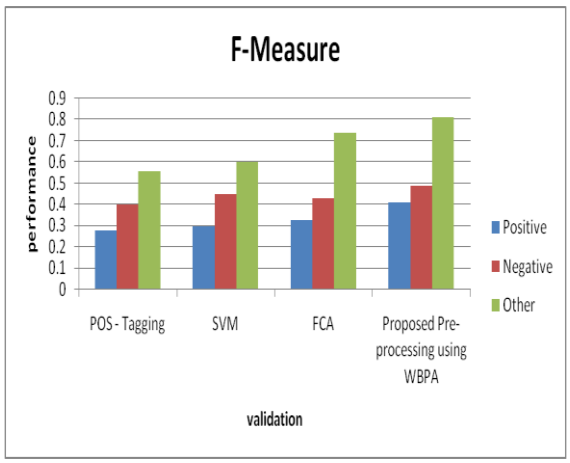

Figure 5.Comparison chart of F-Measure

In Figure 5 illustrates the comparison chart of F-Measure demonstrates the existing1, existing 2 (POS-Tagging, SVM, and FCA) and proposed Pre-processing using WBPA. X axis denote the validation and y axis denotes the performance values in F-Measure. The existing algorithm values start from 0.28 to 0.56, 0.30 to 0.60, 0.33 to 0.74 and proposed Pre-processing using WBPA values starts from 0.41 to 0.81

6. CONCLUSION

In recent years, there has been a considerable rise in social media data which proves that they are a vast amount of data that could be utilized for the decision-making process. Nonetheless, those data are unstructured. Text data pre-processing is one of the effective methods in terms of cleaning and making those unstructured data, structured and meaningful. Thusly, this paper proposes a convincing text data pre-processing technique Weight based pre-processing algorithm (WBPA) to deal with yahoo finance dataset. Three different types of text data pre-processing technique (Stemming, Lemmatization and Spelling Correction)

and its effect on sentiment produced. WBPA can be utilized to provide proportional weight between the hashtag and cleaned text combined to obtain sentiment output. Weight Corrected sentiments can be used to map the overall sentiments produced throughout the year to perceive the popularity of products and obtain insight into its overall performance.

REFERENCES

1. E. Araslanov, E. Komotskiy and E. Agbozo, "Assessing the Impact of Text Preprocessing in Sentiment Analysis of Short Social Network Messages in the Russian Language," 2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI), 2020, pp. 1-4, doi: 10.1109/ICDABI51230.2020.9325654.
2. S. Rathor and Y. Prakash, "Application of Machine Learning for Sentiment Analysis of Movies Using IMDB Rating," 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), 2022, pp. 196-199, doi: 10.1109/CSNT54456.2022.9787663.
3. U. Pasupulety, A. Abdullah Anees, S. Anmol and B. R. Mohan, "Predicting Stock Prices using Ensemble Learning and Sentiment Analysis," 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2019, pp. 215-222, doi: 10.1109/AIKE.2019.00045.
4. A. Krouska, C. Troussas and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), 2016, pp. 1-5, doi: 10.1109/IISA.2016.7785373.
5. S. Pradha, M. N. Halgamuge and N. Tran Quoc Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," 2019 11th International Conference on Knowledge and Systems Engineering (KSE), 2019, pp. 1-8, doi: 10.1109/KSE.2019.8919368.
6. Shetty S, Shetty S. Text pre-processing and parts of speech tagging for Kannada language. Journal of Xi'an University of Architecture & Technology. 2020;12(II):1286-91.
7. A. Krouska, C. Troussas and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA), 2016, pp. 1-5, doi: 10.1109/IISA.2016.7785373.
8. G. Alorini, D. B. Rawat and D. Alorini, "Machine Learning Enabled Sentiment Index Estimation Using Social Media Big Data," 2020 SoutheastCon, 2020, pp. 1-6, doi: 10.1109/SoutheastCon44009.2020.9249672.
9. S. Kumar, N. A. Jailani, A. R. Singh and S. Panchal, "Sentiment Analysis on Online Reviews using Machine Learning and NLTK," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 1183-1189, doi: 10.1109/ICOEI53556.2022.9776850.
10. Y. Hong and X. Shao, "Emotional Analysis of Clothing Product Reviews Based on Machine Learning," 2021 3rd International Conference on Applied Machine Learning (ICAML), 2021, pp. 398-401, doi: 10.1109/ICAML54311.2021.00090.
11. C. Pong-Inwong and K. Kaewmak, "Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration," 2016 2nd IEEE International Conference on Computer and Communications (ICCC), 2016, pp. 1222-1225, doi: 10.1109/CompComm.2016.7924899.
12. S. Abuuznien, Z. Abdelmohsin, E. Abdu and I. Amin, "Sentiment Analysis for Sudanese Arabic Dialect Using comparative Supervised Learning approach," 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 2021, pp. 1-6, doi: 10.1109/ICCCEEE49695.2021.9429560.
13. S. C. Harris and V. Kumar, "Identifying Student Difficulty in a Digital Learning Environment," 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), 2018, pp. 199-201, doi: 10.1109/ICALT.2018.00054.
14. W. Niu and L. Wu, "Sentiment Analysis and Contrastive Experiments of Long News Texts," 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2019, pp. 1331-1335, doi: 10.1109/IAEAC47372.2019.8997550.
15. K. AbdulSattar, Q. Obeidat and M. Akour, "Towards harnessing based learning algorithms for tweets sentiment analysis," 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT), 2020, pp. 1-5, doi: 10.1109/3ICT51146.2020.9311990.
16. S. Zirpe and B. Joglekar, "Negation Handling using Stacking Ensemble Method," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), 2017, pp. 1-5, doi: 10.1109/ICCUBEA.2017.8463946.