Turkish Journal of Computer and Mathematics Education Vol.12 No. 11 (2021), 2900- 2908
Research Article

An Analytical Study Of Feature Extraction Techniques For Student Sentiment Analysis

Mrs. Latika Tamrakar[a], Dr.Padmavati Shrivastava[b] and Dr. S. M. Ghosh[c]

[a] Ph.D. Research Scholar, CSVTU, Bhilai, C.G.
[B]Assistant Professor, BIT, Raipur, C.G.
[C]Professor, Dept. of CSE, CVRU, Bilaspur, C.G.
Email:[a]latika.tamrakar@gmail.com, [b]padmavati.shrivastava@yahoo.co.in, [c]samghosh06@rediffmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract: Student satisfaction is an important factor in the Web-Based Learning System(WBLS). Hence feedback of the Students plays a vital role in the measurement of the effectiveness of any WBLS. The Analysis of feedback or comments is known as Sentiment Analysis (SA) or Opinion mining.SA is the application of NLP used to identify the opinion or emotions behind the comments. Sentiment analysis is a text classification tool that focuses on the polarity of the text (positive, negative, neutral) also emotions (happy, sad, angry)., Classification can be binary (positive or negative) or multi-class. In This paper, we applied two types of Feature Extraction Technique (FETs) namely Count Vector (CV) or Bag of Word (BoW) and Term Frequency and Inverse Document Frequency (TF-IDF).Also presented a comparative analysis of the performance of the machine learning algorithms like Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes(NB), Decision Tree (DT) over Web-based learning models to classify the Student Feedback Dataset (SFD), emphasis is given on the sentiments present in the feedback of the students.

Keywords: Sentiment Analysis(SA), Web-Based Learning System (WBLS), Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Bag of Word (BoW), and Term Frequency Inverse Document Frequency (TF-IDF), Student Feedback Dataset(SFD).

1. Introduction

Learning circumstances are nowadays getting complex, and students have to take additional Accountability for their learning. In the current era, computer-based learning is playing an important role. Along with this, the internet has brought a huge revolution to Web-Based Learning System (WBLS). The WBLS is gaining popularity rapidly, since its early commencement in the 1960s, online education has been repetitively panned for its superficial absence of quality control, particularly the insufficiency of subject-specific teachers, so it's prime important that how much students are satisfied with learning content. In this paper, we will study that how the performance of Web-Based Learning System can be measured with the help of student's feedback in the form of text review. In this study, we generated the dataset by the student's reviews or feedbacks through the web portal http://elearningit.in. The data was collected for the six months, the name of the dataset is the Student Feedback Dataset(SFD). The SFD data set is a text-based dataset. Data pre-processing and cleaning is a challenging task in the text as it is very unstructured so first, it is very essential to prepare the dataset to apply in Machine Learning Algorithms[1], Text and Data Mining (TDM) [2] is an interdisciplinary subfield of data mining and Web-Mining (WM) and measurements with a general objective to extract useful data from a data set and change the data into a conceivable structure for additional utilization. The ML and DM are strongly co-related to each other. In this work, we first pre-processed the text data like cleaned white spaces, numbers, punctuations, stop words, etc. We also used lemmatization and spelling correction. After that, the feature extraction technique is applied. The main task of the Feature Extraction Technique (FET) is to reduce the number of features in a dataset by creating new features from the existing ones and then discarding the original features [1], [3]. In this paper, we Proposed FETs like BoW and TF-IDF. Preprocessing of the text dataset is the first important step for TDM [4]. In this paper, the preprocessed text is converted into the feature vector using techniques like BoW and TF-IDF. The ML techniques namely LR, NB, SVM, and DT are used for the classification of the SFD dataset. The classification performance is compared for the uncleaned SFD Dataset and then the preprocessed SFD dataset for the FETs BoW and TF-IDF. The outcomes are compared in terms of Accuracy, Sensitivity, Specificity, and F1-Score. The word cloud is also used for the frequency analysis of SFD datasets.

2. Literature Survey

KhinZezawar Aung, Nyein NyeinMyo [5] proposed the level of teaching evaluation method based on the lexicon-based approach. This method analyzes the students' feedback comments to strongly negative, or moderately negative, or weakly negative, or strongly positive, or moderately positive, or a weakly positive or neutral category using two lexicons. A heuristic technique is used to calculate the semantic orientation score of combining words for automated students' feedback comments analysis.

Krenare Pireva1, Ali Shariq Imran, FisnikDalipi [6] facial recordings are analyzed to find seven emotional engagement attributes and three sentiment engagement attributes using facial expression software. The author also

proposed some recommendations based on extensive comparison of features among different LMS that will provide better content personalization and customization, thereby improving learning outcomes.

Mohammed Atif [7] author presented an enhanced framework for sentiment analysis using the student's responses and preprocessed data is applied to the classifier model to classify the document whether positive or negative sentiment. This framework showed 0.8 accuracy with 4 grams.

B. Vamshi Krishna, Ajeet Kumar Pandey, and A. P. Siva Kumar [8] proposed a model to analyze user opinions and reviews posted on social media websites. The proposed model uses machine learning techniques and a fuzzy approach for opinion mining and classification of sentiment on textual reviews, Support vector machines (SVM) and Maximum entropy are used for sentiment classification purpose. SVM, F-Score is 0.36

Y. Wang, J. Zhang [9] presented an automatic keyword extraction method based on a bi-directional long short-memory (LSTM) recurrent neural network (RNN). In bi-directional LSTM (BLSTM) network contains two parallel layers that propagate both forward and backward, thus allowing it to obtain information on the sequential series from both the past and future. Each forward or backward layer functions in a similar way to a regular LSTM, Accuracy is 93%.

Ngoc Phuong Chau, Viet Anh Phan, Minh Le Nguyen [10], proposed a model which combines deep learning and sub-tree mining to resolve sentiment classification problem. The association between all words in a sentence and all sentences in a document is captured by LSTM and GRNN, respectively. A document sentiment classification experiment is conducted on a multi-domain sentiment dataset. The elimination of outliers improved Accuracy from 0.14% - 6.93% with LSTM + GRNN model.

Mazen El-Masri, Nabeela Altrabsheha, HanadyMansourb, Allan Ramsay [11] presented a tool that applies sentiment analysis to Arabic text tweets using different parameters. The experiments showed that the Naive Bayes machine-learning approach is the most accurate in predicting topic polarity. Compared between the performance of the lexicon-based method and machine-learning method using Naïve Base and SVM.

3. Backgrounds of Study

The success of outcome-based learning is completely dependent on, the major factor that is student satisfaction, there is a traditional way to attain student satisfaction that we can rate the Learning Management System (LMS) course by using a Likert scale. Many researchers use this technique to collect respondent's opinions. Though this method is very effective but in the present time students are more open to social media platforms and web-based learning is gaining popularity day by day. Now the researchers are giving more attention to the text-based review collected through various sources such as Twitter, Learning portals, and other web-based services. For the analysis of student opinion in WBLS emphasis is given on the comments collected from the students, and machine learning algorithm is applied for the classification of student text review.

4. Proposed Framework

Figure 1 shows the proposed framework for the research work. The data collection for the student feedback dataset, pre-processing of the data set, reduced data set with Feature Extraction Technique, the partition of data set into training and testing dataset, and comparison of the performance of classifier models.

In the first section, we have collected text data in the form of feedback and named the dataset as SFD dataset. Then applied the text pre-processing techniques like removing numbers, punctuations, converts all characters into lowercase, Tokenization, removing stop words, and lemmatization to eliminate the noise and inconsistent data and prepare the smooth dataset. Then we used two different kinds of FET methods like BoW and TF-IDF to extract the features from the SFD dataset. The data partition technique Hold-Out is used to divide the data set into training and testing datasets (70% - training and 30% - testing).In the last section, we compared the classifiers in terms of parameters like Accuracy, Sensitivity, Specificity, F1-Score, and also shown Word Cloud for frequency analysis.

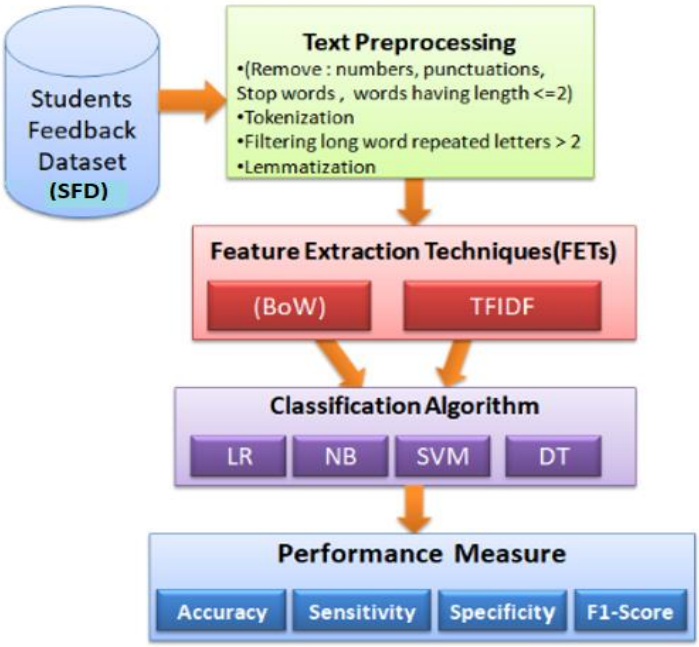

Figure-1. Proposed Frameworks

4.1. About Dataset

In this paper, we used the student feedback dataset (SFD). The student feedbacks are collected from the web portal – http://elaerningit.in developed for students learning. The data collected from January 2020 to June 2020. In this web portal students get enrolled in any of the courses available in the web portal and after completion of the course students give their feedback about the course, content, and the LMS. If they are satisfied they provide positive comments and if not satisfied then provide negative comments. The original dataset consists of 549 comments of students. Based on the comments dataset can be divided into two class labels Positive comments and Negative comments. The positive comments are labeled as 1 class name label and for negative comments, the value of the class label is 0.

4.2 Text Preprocessing

The feedbacks taken by the students are in form of natural language i.e. in the English Language as the machine learning model doesn't understand input in form of the text so we first need to convert it into a form that the machine learning model can understand. Before converting the text into numbers or vectors the very first step that we need to follow is preparing data to be sent in the model. As there are many challenges involved with text data, it contains lots of noises as people usage punctuations, slang, emoticons, and spelling mistakes are also there. E.g. they use sorrryyyy ,veryyyy, gr8, sooooo much, this kind of word which machine cannot make sense out of it and some word which is most frequently used like I, you, he, she, is, am, the, these kind words called stop words which don't carry any emotion so it is always good to remove these words to increase the accuracy of the model.

The preprocessing steps which are carried in this SFD dataset are:

i. Removing Numbers and Punctuations
ii. Converts all Characters into Lowercase
iii. Tokenization
iv. Removing Stop Words
v. Lemmatization
vi. Removing the words having Length <=2
vii. Filtering long word repeated letters > 2
viii. Spelling correction

4.3 Feature Extraction Techniques

Feature Extraction Techniques (FETs) have a significant role in text dataset classification as they affect the accuracy of text classification. It depends on the vector space model, (VSM), in which a text is converted into N-dimensional space [12]. For converting the text into features some feature extraction techniques need to be applied. Here we used methods like Countvectorizer (bag-of-words) or TF-IDF to create features, we take into account all

the tokens occurring in the dataset and these tokens determine the dimensions which are nothing but the number of features.

In this paper we used two of the most basic and ubiquitously used formats:

4.4.1 Count Vector (Bag of words)

The Bag of Words (BoW) model is the simplest form of text representation in numbers. A bag of words is a representation of text that describes the occurrence of words within a document. This model is used to convert the text into a bag of words, which keeps account of the total occurrences of the most frequently used words. The model is only concerned with whether known words occur in the document, not wherein the document.[13] [14]The bag-of-words model is most commonly used in methods of document classification where the frequency of eac word is used as a feature for training a classifier.

4.4.2. TF-IDF Vector

Term Frequency–Inverse Document Frequency (TF-IDF) is a statistical method that reveals that a word is how significant to a document in a collection or corpus. The TF-IDF is frequently utilized as a weighting factor in the text mining method. The value of TF-IDF increases proportionally to the number of times a word appears in the document but is counteracting by the frequency of the word in the corpus[15]. Term Frequency- Inverse Document Frequency (TF-IDF) methods were quite popular for a long time, before more advanced techniques like Word2Vec or Universal Sentence Encoder. Term Frequency of a particular word is calculated as the number of times a word occurs in a document to the total number of words in the document. IDF is used to calculate the importance of a term. It shows how important a word is to a document.[16]

4.4 Text Classification

Text classification is a technique to systematically classify text objects (document or sentence) in a fixed category i.e. ordering the dataset into decided text classes. Text Classification in supervised learning comprises two stages: training and testing. In the training phase, a classifier was trained using the training data set, and the trained model tested using the testing data set [17]. There are four classification methods used in this work for the classification of the SFD dataset.

There are four classification methods used in this work for the classification of the SFD dataset.

- Naïve Bayes:

Naïve Bayes is a simple method that uses all the attributes and permits them to contribute to take the decision, all the features as equally important and independent of each other.Naïve Bayes text classification method is based on the Bayesian theorem, Using Prior probabilities to classify new text. The Naive Bayes (NB) classifier used in this study is the Multinomial Naive Bayes classifier(MNNB) [18].

- Decision Tree:

A “divide-and-conquer” approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a Decision Tree[DT][19] In this study, we have used CART as a DT. CART is a non-restrictive DT method used to build model either classification or regression trees, based on whether the dependent variable is categorical or numeric. It constructs a binary DT by isolating the record at each node, according to a function of a single attribute [3].

- **Logistic Regression:**

The logistic regression is an overall measurable model that shows the probability of a specific class for example, great/terrible, pass/fail. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Since it predicts the probability, its output values lie between 0 and 1. The logistic regression model itself just models the probability of yield regarding input and doesn't perform statistical classification, however, it very well may be utilized to make a classifier. This is a typical method to make a binary classifier. [21].

- **SVM:**

A Support Vector Machine (SVM) [22], [23] is another strategy for the classification of both direct and nonlinear text data. It depends on the idea of decision planes that describe decision limits. A decision plane isolates between a lot of items having different class participation. An SVM is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model set of labeled training data for each category, they're able to categorize new text. Particularly good for very sparse data in very high dimensional spaces.

4.5 Performance Evaluation

In this work, a confusion matrix of 2x2 matrix is used for evaluating the performance of the classifier models. Here 2 is no of target classes. The confusion matrix contains four promising groups True positive (TP), False Positive (FP), True Negative (TN), False Positive (FN) as shown in table 1.

	Hypothesized class or predicted class		
Actual class or Observation		Class **+Ve**	Class **–Ve**
	Actual **+Ve**	TP(**+Ve, +Ve)**	FN(**-Ve, +Ve)**
	Actual **- Ve**	FP(**-Ve, +Ve)**	TN(**-Ve, - Ve)**

Table 1. Confusion Matrix

The different parameters used for classification performance evaluation are shown in equations (1) to (4).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(1)}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{..... (2)}$$

$$\text{Specifidity} = \frac{TN}{TN+FP} \quad \text{.....(3)}$$

$$\text{F1-Score} = \frac{2TP}{(2TP+FP+FN)} \text{.....(4)}$$

5. Result and Discussion

The experiment conducted for the analysis of this research work is done with Python programming version 3.7.3. using Jupyter Notebook under Anaconda 3. The result and discussion sections are divided into four different sections according to work nature and outcomes. In this paper, we have used the following terms as Unclean Student Feedback dataset (USFD) for raw SFD dataset and for pre-processed or Cleaned SFD or Normalized SFD dataset (NSFD) respectively. Then Data Partition Technique (DPT) is applied to the SFD dataset using the hold-out method. The dataset is divided into 70% for Training and 30 % for Testing in models

5.1 Preprocessing of SFD dataset

Working with text generally involves converting it into a format that our model can understand, which are mostly numbers. In the initial stage of the text data pre- preprocessing, we applied different kinds of pre-processing steps such as removed numbers and punctuations, converted all uppercase to lowercase, tokenization, removed stop words, lemmatization, removed the words length 2 or less, converted list to strings. The SFD dataset can be categorized into two categories according to the actual class labels, Positive and Negative Reviews. In this work, we used the SFD dataset which contains a total of 549 comments, out of which 446 are positive and 103 are negative comments. The SFD Database contains the feedback given by students for six months.

5.2. Feature Extraction Technique

As machine learning algorithms cannot work on the raw text directly, in this work we applied BoW and TF-IDF techniques as FETs.Here CountVectorizer of Sklearn library of python Language is used to create count vectors from the text. Count Vectorization involves counting the number of occurrences each word appears in a document. After counting the words, it forms a Sparse matrix. A sparse matrix is a matrix that has very few non-zero elements. The count of words matrix creates the data frame. In the Data frame, each row represents the given text in 'data and columns represent the unique words from the given string of lists, and values shown in the Data Frame table are the occurrence of words. In the same way, TF-IDF is also implemented in Python language using the Sklearn library. While exploring the data frame in TF-IDF FETs, we get a numeric value for each word in the corpus, which is the TFIDF score of that word.

5.3. Machine Learning Classifiers (MLC) performance

The confusion matrix obtained by proposed ML-C algorithms(LR, NB, SVM, and DT) is shown in Table

FETs →	**BOW**					**TF-IDF**			
ML-C		**TP**	**FN**	**FP**	**TN**	**TP**	**FN**	**FP**	**TN**
LR	**USFD**	8	25	1	131	4	29	0	132
	NSFD	9	24	5	127	4	29	1	131
NB	**USFD**	17	16	17	115	4	29	1	131
	NSFD	16	17	13	119	5	28	1	131
SVM	**USFD**	16	13	9	127	14	19	3	129
	NSFD	20	13	8	124	15	18	3	129
DT	**USFD**	24	9	39	93	25	8	28	104
	NSFD	27	6	30	102	26	7	23	109

Table 2: Confusion Matrix of Machine Learning Classifier (ML-C)

The confusion matrix received by the ML-C methods is shown in Table-2 with the holdout method (Training 70%, - Testing 30%) in the case of the USFD and NSFD datasets. The TP achieved the highest value by DT for both BoW and TF-IDF FETs using the NSFD dataset. TN is acquired best by LR with the TF-IDF for both USFD and NSFD datasets. The FP is achieved maximum by DT with the FET BoW with the USFD dataset. The FN is maximum for TF-IDF by LR for both USFD and NSFD datasets.

ML-C		Accuracy	Sensitivity	Specificity	F1-Score
LR	USFD	84.24	24.24	99.24	38.09
	NSFD	82.42	27.27	96.21	38.30
NB	USFD	80.00	51.52	87.12	50.75
	NSFD	81.82	48.48	90.15	51.61
SVM	USFD	86.67	51.52	96.21	61.82
	NSFD	87.27	60.61	93.94	38.09
DT	USFD	70.91	72.73	70.45	50
	NSFD	78.18	81.82	77.27	60

Table 3. Performance of ML Classifiers with Bow in SFD dataset

Table-3 shows the performance of ML-C algorithms using BoW FET in the SFD dataset. In this table the Accuracy is shown for two cases - The uncleaned dataset i.e USFD and for the Normalized dataset i.e. NSFD dataset. From the above dataset, we get 86.67 % Accuracy for USFD, while 87.27% Accuracy for NSFD. Hence preprocessed dataset improved the performance.

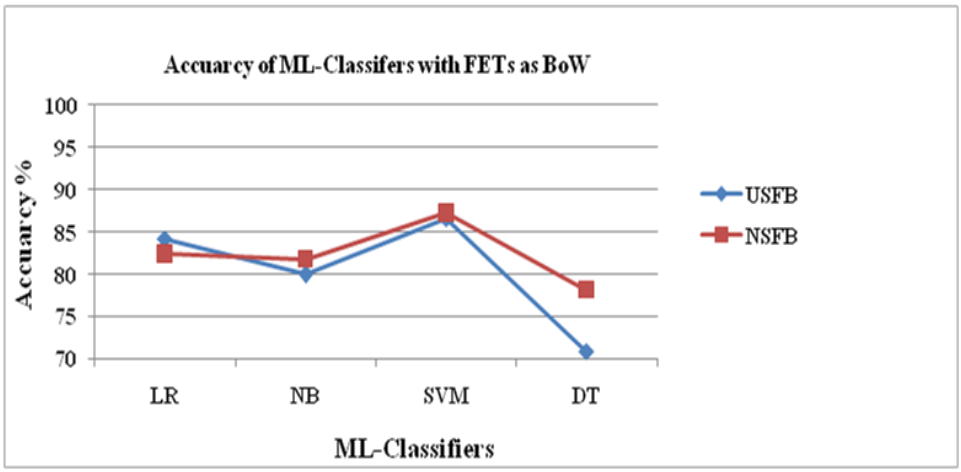

Figure 3: Comparison of Proposed models ML-C in BoW

The comparison of the model performances (LR, NB, SVM, and DT models) is shown in Figure -3. The outcome of the classification models Accuracy graph 3 of proposed ML-C models obtained better accuracy with the cleaned

NSFD as compared to the USFD dataset. The classification result shows that the SVM model is better than other models used for the analysis in the case of FET as Bow.

ML-C		Accuracy	Sensitivity	Specificity	F1-Score
LR	USFD	82.42	12.12	100	21.43
	NSFD	81.82	12.12	99.24	20.87
NB	USFD	81.82	12.12	99.24	20.87
	NSFD	82.42	15.15	99.24	25.41
SVM	USFD	86.67	42.42	97.73	55.55
	NSFD	87.27	45.45	97.73	58.36
DT	USFD	78.18	75.76	78.79	58.08
	NSFD	81.82	78.79	82.58	63.44

Table 4 . Performance of ML-C classifiers with TF-IDF in SFD dataset

Table 4. shows the Performance of ML-C classifiers with TF-IDF in the SFD dataset. The ML classifiers Accuracy are shown here for USFD and NSFD datasets. The SVM classifier's Accuracy is 86.67% with USFD datasets and SVM classifiers obtained the Accuracy of 87.27% with preprocessed NSFD. So it shows that when applied preprocessing on the dataset, the performance of models is improved.

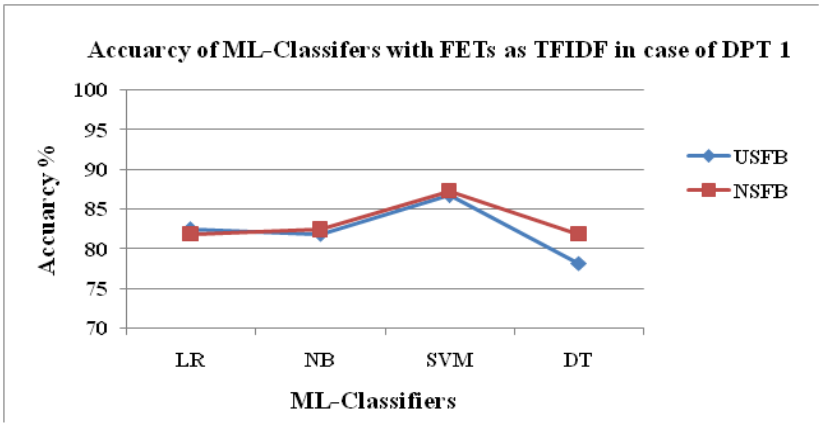

Figure 4. Comparison of Proposed models ML-C in TF-IDF

Figure 4 shows the comparison of the model performances (especially LR, NB, SVM, and DT) models. The outcome of the classification models Accuracy graph 4 of proposed ML-C models obtained the best accuracy with the cleaned NSFD as compared to the USFD dataset. The classification result shows that the performance of the SVM classifier model is better than other models used for the analysis for FET as TF-IDF.

5.4. Word Cloud Visual Representation of SFD dataset

A word cloud is a collection or cluster of words shown in different sizes. It is a very important technique to represent the student comments that have big value or less value in WBLMS. The biggest word shows the highest frequency of a word in comments. There are multiple ways of visualizations in the form of word clouds. Sometimes, the quickest way to understand the context of the text data is by using a word cloud of the top 100-200 words. Here we created a Word cloud for our most frequently used words in the SFD dataset.

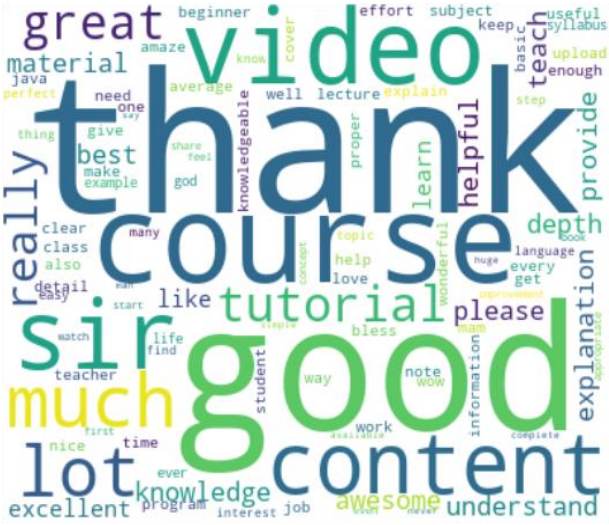

Figure 5. The word cloud of SFD dataset.

Figure 5. represents the word cloud of the SFD dataset. From the figure, we get the idea that the highest frequency word is thanks in the SFD dataset. Then the good is the second frequently used word. With the help of the above word cloud, we can conclude that in received feedback we have more no of positive words like thanks, good, great, helpful, etc. It means the WBLMS system is useful for the students in learning and understanding the concept. The content used in the WBLMS system gives a satisfactory result for the review analysis of student feedback.

6. Conclusion

In this paper, an SFD sentiment classification model is proposed to identify the student's sentiment as Positive and Negative by the feedback given by them for WBLMS. The feedback collected in the form of text WBLMS forms a raw dataset. These comments show the strength and weaknesses of the WBLMS. The Obtained raw SFD dataset is first pre-processed. After that, we used two feature extraction techniques Bow and TD-IDF to covert the raw text into feature vectors. The Bag of Words (BOW) converts the collection of text documents to a matrix of feature vector counts that gives the no of occurrences a word appears in an SFD dataset. The TF-IDF method represents that a word is how significant to a document. Based on the SFD dataset, the comments are classified as positive and negative. The model is examined for FETs, BoW & TF-IDF to compare the performance. The result shows that the performance of the model is improved when preprocessing is applied in the uncleaned SFD dataset. The experimental results show that the SVM algorithm performed the best with 87.27 % in the case of both BoW and TF-IDF FETs.

References

1. J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Third. Elsevier, 2012. A. Pujari, Data mining techniques, Third. University press, 2013.
2. K. Shrivas, S. K. Sahu, and H. S. Hota, "Classification of Chronic Kidney Disease with Proposed Union Based Feature Selection Technique," SSRN Electron. J., no. 2007, pp. 503–507, 2018, doi: 10.2139/ssrn.3168581.
3. S. Vijayarani, J. Ilamathi, and Nithya ., "Preprocessing Techniques for Text Mining," Int. J. Comput. Sci. Commun. Networks, vol. 5, no. 1, pp. 7–16, 2015.
4. K. Z. Aung and N. N. Myo, "Sentiment Analysis of Students ' Comment Using Lexicon Based Approach," in IEE Computer Society, 2017, pp. 149–154.
5. K. Pireva, A. S. Imran, and F. Dalipi, "User behaviour Analysis on LMSs and MOOCs," 2015, no. March 2019, doi: 10.1109/IC3e.2015.7403480.
6. M. Atif, "An Enhanced Framework for Sentiment Analysis of Students ' Surveys : Arab Open University Business Program Courses Case Study," Bus. Econ. J., vol. 9, no. 1, pp. 9–11, 2018, doi: 10.4172/2151-6219.1000337.
7. V. Krishna, A. K. Pandey, and A. P. S. Kumar, "Feature Based Opinion Mining and Sentiment Analysis Using Fuzzy Logic," in Cognitive Science and Artificial Intelligence, 2018, pp. 79–89.

8. Y. Wang and J. Zhang, "Keyword Extraction from Online Product Reviews Based on Bi-Directional LSTM Recurrent Neural Network," in Proceedings of the 2017 IEEE IEEM, 2017, pp. 2241–2245.
9. N. P. Chau, V. A. Phan, and M. Le Nguyen, "Deep Learning and Sub-Tree Mining for Document Level Sentiment Classification," in Eighth International Conference on Knowledge and Systems Engineering (KSE), 2016, pp. 268–273.
10. M. El-Masri, N. Altrabsheh, H. Mansour, and A. Ramsay, "A web-based tool for Arabic sentiment analysis," Procedia Comput. Sci., vol. 117, pp. 38–45, 2017, doi: 10.1016/j.procs.2017.10.092.
11. H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," Eurasip J. Wirel. Commun. Netw., vol. 2017, no. 1, pp. 1–12, 2017, doi: 10.1186/s13638-017-0993-1.
12. R. Zhao and K. Mao, "Fuzzy Bag-of-Words Model for Document Representation," IEEE Trans. Fuzzy Syst., vol. 26, no. 2, pp. 794–804, 2018, doi: 10.1109/TFUZZ.2017.2690222.
13. T. Walkowiak, S. Datko, and H. Maciejewski, Bag-of-words, bag-of-topics and word-to-vec based subject classification of text documents in Polish - A comparative study, vol. 761. Springer International Publishing, 2019.
14. S. Kannan et al., "Preprocessing Techniques for Text Mining," Int. J. Comput. Sci. Commun. Networks, vol. 5, no. 1, pp. 7–16, 2015.
15. R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," Procedia Comput. Sci., vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.
16. S. K. Sahu and P. K. Chandrakar, "Classification of Chronic Kidney Disease with Genetic Search Intersection Based Feature," in Advances in Intelligent Systems and Computing 1122, vol. 1, 2020, pp. 11–21.
17. Jurafsky and J. H. Martin, "Naive Bayes and Sentiment Classification," in Speech and Language Processing, 2019, p. Chapter 4.
18. H. Witten and E. Frank, Datamining. Practical Machine Learning Tools and Technicals with Java Implementations., Second. Elsevier, 2004.
19. J. S. Cramer, "The origins of logistic regression," 2002.
20. B. U. Park, L. Simar, and V. Zelenyuk, "Nonparametric estimation of dynamic discrete choice models for time series data," Comput. Stat. Data Anal., vol. 108, pp. 97–120, 2017, doi: 10.1016/j.csda.2016.10.024.
21. Vapnik V, Statistical Learning Theory. New York: John Wiley and Sons, 1998.
22. Cortes and V. Vapnik, "SUPPORT-VECTOR NETWORKS," Kluwer Acad. Publ. Boston. Manuf. Netherlands., vol. 297, no. 973, pp. 273–297, 1995.