International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hybrid Feature Selection Techniques for Aspect based Sentiment Classification using Supervised Machine Learning

Mr. Ganesh N. Jorvekar[1] Dr. Tripti Arjariya[2] Prof. (Dr.) Mohit Gangwar[3]

Submitted: 20/10/2022 **Revised:** 25/12/2022 **Accepted**: 24/01/2023

Abstract: Sentiment Analysis comprises a variety of tasks, including subjectivity detection, polarity detection, sentiment magnitude detection, and emotion type recognition. The text is assessed as subjective in the subtask of subjectivity detection. Words and phrases' subjective nature vary based on their context; an objective text may include subjective observations. For example, news reports include quotations from people's viewpoints. In this paper, we proposed an aspect-based sentiment classification using machine learning techniques on a customer review dataset. The real-time customer review dataset has been considered for the identification of aspects and later for the classification of sentiment. The various feature extraction and selection techniques have been carried out for unique module building, while numerous machine learning (ML) classification algorithms have been used for the classification of aspects as well as the sentiment. Several machine learning (ML) classification techniques such as artificial neural network (ANN), RF (Random forest), Naïve Bayes (NB) and SVM (support vector machine) have used for both classifications. The three feature extraction techniques have been used during the implementation such as TF-IDF, Bigram as well as NLP features. In an extensive experimental analysis, SVM obtains better results with NLP features over other machine learning classifiers.

Keywords: *Supervised machine learning, aspect identification, sentiment classification, Natural language processing, customer review dataset*

1. Introduction

The widespread use of e-commerce and social media in the 21st century has led to the generation of massive unstructured data which is publicly accessible. This unstructured data primarily consists of user reviews regarding products and services as well as opinions and emotions on social and political issues. Text, photos, music, videos, and emojis are all examples of unstructured information. The automated analysis of this unstructured data is of enormous importance to successful business organizations and governments. This led to the emergence of Affective Computing and Sentiment Analysis as a specific discipline within Artificial Intelligence [1].

ABSA (Aspect Based Sentiment Analysis) methods use an input text collection which describes a particular object. The systems try to identify the object's primary characteristics and calculate the average emotion of the texts per characteristic. Although many ABSA methods have been suggested, most of which are development models [1]. There is no defined task breakdown for ABSA and no recognized assessment metrics for the embedding process of ABSA systems. This work introduces a new task involving sentiment called aspect identification and classification techniques, which consists of three subtopics: aspect term retrieval, aspect term polarization estimation, and classification techniques. The first subtask identifies single- and inter phrases identifying characteristics of the object under discussion; these expressions are referred to as aspect phrases from here on out. The second subtask calculates the average sentiments per aspect word or clusters multiple aspect phrases, and the third subtask performs sentiments classification using ML (machine learning) approaches.

During this work, benchmarking datasets for various objects were created for each of the subtasks mentioned earlier. New assessment techniques are presented for each process step, which is more suitable than primary evaluation methods. The dissertation also offers new designs for each subtask, demonstrating that the new techniques are superior or at least similar to state-of-the-art approaches empirically on the built standard datasets. For the feature word extracting subtask which is authorized by a global challenge, novel artificial datasets were produced. It was also demonstrated that there is

[1] *Ph.D. Research Scholar, Department of Computer Science and Engineering, Bhabha University, Bhopal, Madhya Pradesh, India*
[2] *Head, Department of Computer Science and Engineering, Bhabha University, Bhopal, Madhya Pradesh, India*
[3] *Director (Alumni Cell),B.N. College of Engineering and Technology, Lucknow*
[1] *ganesh.jorvekar83@gmail.com*, [2] *drtripti.beri@gmail.com* ,
[3] *mohitgangwar@gmail.com*

acceptable concordance when experienced judges are requested to annotate aspect words in documents. Consumers or users commonly contribute their ideas, experiences, or opinions about any product or news, as seen by the recent surge in the collecting of reviews, comments, or feelings from a range of online marketing and social networking sites. As a consequence, new users, manufacturers, and sales managers may benefit from these evaluations. They may get extensive information on the product's quality, which will help them decide whether to purchase, create, or sell it. Similarly, when it comes to movies, people remark on the quality of the film. The issue using these assessments is that they are usually presented in text format, which requires extensive processing to extract any useful information. Sentiment Analysis (SA) satisfies this desire by analysing and categorising customer feedback depending on their demands.

2. Literature Survey

In this section we describe a work done by existing researchers for aspect detection and sentiment classification using ML (machine learning) methods. It is also advised to compare classifiers pairwise using experimental assessment measures. The Wilcoxon test is a good nonparametric test here. The Wilcoxon test assumes that the median difference between the two experimental procedures is zero. The chance of rejecting a valid null hypothesis is called the significance level in hypothesis testing. In empirical investigations, p value of 0.05 is used [2]. The observed p-value is termed as p-value. The null hypothesis is rejected if indeed the p value is equal to or less than 0.05, indicating that two deep learning algorithms perform meaningful but differently. Unsupervised, semi-supervised, and supervised techniques are available for aspect level sentiment categorization. Surveillance methods include corpus and lexical methodologies [3]. The corpus-based strategy uses massive domain-specific corpora to provide relevant information, requiring a lot of human work, training, and data. The lexicon-based technique uses external sentiment lexicons. These strategies are reliant on the quality of the information base and often suffer from word limitations.

ULMFiT was established by Howard and Ruder in 2018 as element of the fast.ai system [4]. ULMFiT is an information extraction method based on transfer learning. To training model, this approach just requires a little amount of information. ULMFiT is also global, which means it may be applied to almost any dataset or document using only a single architecture, such as AWD-LSTM. This method's versatility means it may be used for any purpose without requiring unique engineering. The language model relies on RNN (recurrent neural networks) as well as AWD-LSTM. This strategy emphasizes familiarising yourself with common phrases and word connections [5]. The first layer is called embedding (black), after the term embedding. Modern models utilize the same embedding layer for RNN layers. The RNN layers are the implicit feedback layers. In the hidden layers, phrases are substituted by identical embedding vectors, which are subsequently transmitted to the RNN layer [6].

The primary idea of SVM (Support vector machine) is to classify data independently using a hyperplane to maximize the border between them. Support vector machine (SVM) is often employed in statistical learning theory (SLT) based on classification systems. SLT has been used in computer vision, speech recognition, and bioengineering. SVM is commonly utilized in problem classification because of its high classification accuracy and performance [7]. SVM is a powerful ML technique for sentiment analysis [8]. SVM is a popular classifier because it may provide greater generalization performance when the input feature space's dimension embedding is high. The goal of SVM is to find the best classification function for the training data. The SVM classification creates a hyperplane for excluding positives from a set of negatives. Support vector machine seeks to categorize points of data onto a hyperplane. SVM optimizes the support vector margin to separate all classes [9]. In the actual world, SVM is utilized to address issues like intrusion detection, image processing, text categorization, etc. It was created to categorize binary classes. The examples later included many classes. The binary classification problem divides a collection of cases into two groups and determines whether or not they contain certain qualities. Real-world tasks that include binary classification include object recognition, figuring out connections to a specific class of the object, face identification, and ID (intrusion detection). The SVM algorithm's mathematical foundation is shown in two binary classification tasks: linear divisible and non-linear divisible. Any hyperplane that splits the two categories represented by the training data is non-linear divisible. The well-known question is how to pick the appropriate hyperplane for precision performance on test data. The optimum option is to use the support vectors' boundary, which separates the negative as well as positive points in the train dataset. A hyper - plane that splits the 2 groups evenly is the strongest.

LSTM is an RNN extension. Gradient vanishing was suggested for vanilla RNN training. With its unique memory system, it outperforms standard RNNs. The memory method allows the network to capture long-term relationships, such as the LSTM approach in [10], which reduced dimensionality and performed very well in opinion categorization. Reducing input functions is critical for machine learning sentiment categorization. For improved categorization with scalability, the proposed strategy may be promising. The suggested technique is useful for big dataset applications such as sentiment identification in evaluation of service and product. RNN, Tarasov [11] is used as a long-short-term approach. To evaluate the collecting outcomes, simple RNN, logistic regression (LR), bidirectional recurrent neural network, and bidirectional long-term memory are being used. Among all recurrent neural network models, bidirectional deep LSTM with multiple hidden layers is outperformed. To analyse the general model text data, a system training parameter could be a vector representation of single words. To reproduce particular variables, we may start single sentences using random vector representations. Tai et al. [12] used LSTM to identify amorous terms from film reviews and estimate somatic sentence pair estimates. BPTT (Back-propagation through time) is used for categorization of text and language modelling. This method divides text into fixed size chunks. The model starts each portion with the previous portion's ultimate state. For the ultimate prediction, the hidden layer of the gradients is returned to the batches. The most popular practical paradigm is variable-length back-propagation [13]. Daniel Langkilde utilized linear SVM to classify airline tweets in 2017. The accuracy rate was 78percent, and the researcher recommended that additional tweaking or a more sophisticated method may improve the score [14]. Wesley Liao used Fast.ai's system for ULMFiT to evaluate the Twitter United States Airline sentimental dataset in 2019.

Saikat Bose et. al. [15] advocates for the use of a novel data security protocol to verify the appointment of candidates for service. The process began with the private information being obfuscated in the e-initial mail's section for each area on the server run by the commission. Circular orientation of private share pieces and their hosted matrix intervals are determined by hash operations. The same hash operations and public sharing are used to verify any digitally signed letters that are downloaded from the designated location. On-the-spot fingerprints are hidden using identical concealment techniques in two sections for each section of the electronic letter. Each region's fourth segment is encrypted using a hash function to protect the copyright signature of the posting location. The commission's server verifies the legitimacy of the appointment and the validity of the candidate's signatures to ensure that the certified electronic letter is sent in its whole to the designated location. The effectiveness of the suggested procedure is established above the previous ways by the improved test findings from broader angles.

3. Proposed System Design

The below Figure 1 depicts the proposed system architecture for aspect detection and sentiment classification. The system's aspect prediction function is evaluated on the restaurant review dataset. This research made use of the customer review dataset. One or more aspect categories are assigned to each review in the dataset. The dataset contains 5000 training reviews, with 20% of the statements comprising more than two aspect categories. It was formerly considered as a single-label issue, with multi-class classification being used to predict a label. The suggested method eliminates this drawback. Multi-label classifiers are used in the suggested approach to predict the class labels of a review. Multi-label classifiers perform better using a two-phase feature selection strategy.

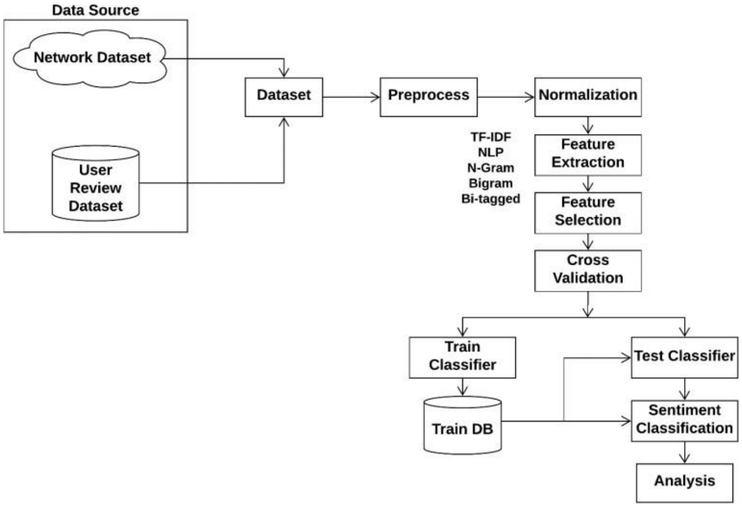

Fig. 1: proposed system design for aspect-based sentiment classification using machine learning techniques

Pre-processing and data filtration

In the preprocessing phase tokenization, for upper to lowercase transformation, stop word filteration, and stem are all applied to each phrase in the training dataset. The stop word dictionary, which can be found at https://gist.github.com/larsyencken/1440509 is utilized. The two most important feature normalization techniques utilized in the preprocessing step are stemming and lemmatization. The stemming technique is used to restore all of the impacted phrases in the text to their original form, also known as stem phrases. For example, the stems 'study' and 'studi' are formed from the words studying and studies. The conversion of all types of words to their basic lemma is the main goal of lemmatization. The words "studying" and "studies," for example, will be transformed to the lemma "study." As a result, lemma features are thought to be more accurate than stemmed features. Lemmas are retrieved as features in this experiment and then subjected to feature selection techniques.

Feature Extraction and Selection

Various feature selection methods are examined in the system, and a hybrid feature selection strategy is suggested. The feature selection methods that have been investigated are as follows:

Term frequency (TF)

The term frequency count is used to pick features in this method. The term frequency of every feature is computed for every aspect class. For feature selection, a threshold is specified. In every aspect classes, features are selected with a word frequency higher than '2'. As a consequence of this, a term frequency matrix for each aspect category is produced. In addition, a compound matrix comprising words and their recurrence counts in all aspect classes is created. A matrix of binary train is produced from this matrix, with '1' representing non-zero term frequency.

TF with Weight calculation

These techniques utilized the weight of every term and it is computed by using equation in this technique (1). The conditional probability of a term is X (t,k), where X t is the overall occurrence count of a term "t" in all aspect classes and X (t,k) is the occurrence count of term "t" in aspect classes "k." The weight of "t" rises if the percentage of occurrence of a word "t" in aspect classes "k" is higher than the other aspect classes. For every aspect classes, a weight threshold is set. To create a binary train matrix, terms (i.e features) with a weight higher than that of threshold are selected.

$$\text{weight}(t) = \frac{X_{t,k}}{\mathrm{X}_t} \qquad (1)$$

Term Frequency with relational features

To enhance the classification preciseness of the classifier, attributes are useful but not redundant. The term frequency matrix produced in I is utilized in this approach. This matrix gives valuable data, but it also provides unnecessary features. To prevent duplication, each feature's correlation is computed in relation to other characteristics in the same aspect category. To calculate correlation, the Pearson correlation coefficient is employed.

$$C0\_weight\ [t_i] = \frac{\mathrm{n}\ (\sum \mathrm{X}[]\mathrm{Y}[]) - (\sum \mathrm{X}[]) * (\sum \mathrm{Y}[])}{\sqrt{[\mathrm{n} \sum \mathrm{X}^2 - (\sum \mathrm{X})^2]}\sqrt{[\mathrm{n} \sum \mathrm{Y}^2 - (\sum \mathrm{Y})^2]}} \qquad (2)$$

Equation (2) is used to compute the correlation of every term with regard to certain other terms, wherein "x[] and y[]" are vectors of term t (i) and t (i+1), correspondingly, representing term frequency with regard to every aspect class. The mean of correlation of word "t" with some other terms in the same aspect class is evaluated. Terms having a correlation value that is less than or equal to 0.85 are picked to form a binary train matrix.

Term frequency with weight calculation

The weighted matrix produced in (ii) is utilized to create a new matrix containing the weight of a word in relation to every aspect class in this method. The correlation of each word with regard to other terms is calculated using Equation (2), where "x[] and y[]" are vectors of term t_i and t_{i+1}, correspondingly, including the weight of term "t" in each aspect class. Lastly, as previously stated, a binary train matrix is produced (iii).

This article makes a contribution by proposing a supervised method for aspect class extraction that chooses important attributes while avoiding duplication by assessing feature correlation. The obtained findings indicate that the weighted word frequency with correlation method has a higher F-score than the other approaches. It is discovered in this study that characteristics chosen using weighted word frequency are not only more significant, but also redundant. By measuring the correlation between characteristics in an aspect category, redundancy is eliminated.

Classification: According to a selection of features, the model training has done with several machine learning classifiers. A similar classification algorithm has been used for module testing. In experimental analysis validation has done with different existing system and proposed classification technique

Algorithm Design

Algorithm 1: Selection of features for aspect identification

```
Input:, Hybrid feature set such as TF-IDF, NLP and N-gram features
Output: Training matrix or rule generation
Step 1 : Fset = All Features of TF
Step 2 :  for each (Fq in Fset)
            Frequency(Fq) ← TF[fq]
            if (Frequency(Fq) ≥ TFDenominator)
            weightList.add(Fq)
 end for
Step 3 : for every (t in weightList)
            aspectTA ← aspect[t]
            Owen_Count ← t.occurances(aspectTA)
            Other_count ← t.occurances(! aspectTA)
    $aspect\_weight(t) = \frac{\mathrm{Own}_{[\mathrm{Count}]}}{\mathrm{Other}_{[\mathrm{count}]} + \mathrm{Own}_{[\mathrm{Count}]}}$
           Update weight for t ← aspect_weight(t)
End for
Generate weighted matrix according to calculated weight for particular term which fulfils the various threshold criteria for each category.
Step 4 : To calculate the correlation coefficient weight for each term using below X and Y matrix, here $X(term)[]$ represents the matrix of current term t with all n aspect index as well as $Y(term)[]$ illustrates next terms t+1 index matrix. Both matrixes should have a same index, before calculation of correlation weight of term t with t+1.
    $X(term)[] = \sum_{p=1}^{n}(.\ aspect[p] \ldots \ldots ..\ aspect[n]\ .\ count\ )$
    $Y(term+1)[] = \sum_{q=1}^{n}(.\ aspect[q] \ldots \ldots ..\ aspect[n]\ .\ count\ )$
Step 5: Apply coefficient correlation on X and Y matrix using below formula
    $C0weight\ [term] = \frac{\mathrm{n}\ (\sum \mathrm{X}[]\mathrm{Y}[]) - (\sum \mathrm{X}[]) * (\sum \mathrm{Y}[])}{\sqrt{[\mathrm{n} \sum \mathrm{X}^2 - (\sum \mathrm{X})^2]}\ \sqrt{[\mathrm{n} \sum \mathrm{Y}^2 - (\sum \mathrm{Y})^2]}}$
Step 6: To select the best feature from existing feature set after correlation, coefficient weight is calculated using given threshold
if ($C0weight\ [term] \le th$)
            FinalFeatureset ←{Term, Aspect}
Step 7: Generate final matrix for each term called correlation coefficient matrix
```

Algorithm 2: Classification Algorithm

```
Input: Weighted coefficient correlation matrix for each selected term M, Test dataset test_data
Output: aspect label prediction for every instance
Step 1: for every t in M
Step 2 : read all index M [j…n] values
Step 3 : if ($M[j] \ge 0.0$)
          Newlable ← Convert label 1 for respective t
          MN[] ← Newlable[j……n]
End for
Step 4 : create updated binary matrix MN[]
Step 5 : input and preprocess test_data
    $Lemmas[] = \sum_{k=0}^{n} input[k\ \ldots .\ n].\{Stopword, Lemmitization\}$
Step 6: for each ($s\ in\ Lemmas$ )
        T[] ← s.split(tokens)
For each (term t in T)
        $f(x) = t\ ||\ \sum_{n=1}^{m}(MN[n].\ values)$ if(exist)
Step 7 : calculate mean for each aspect category based on $f(x)$
Step 8 : calculate belief for all categories
Step 9 : Return highest belief category as aspect for test instance.
End for
End for
```

4. Results and Discussions

The proposed model was developed using Java technology in a Windows environment, and some inbuilt methods were used to identify and extract characteristics. The below unit explore the various experiment with different machine learning classifiers with numerous feature extraction techniques.

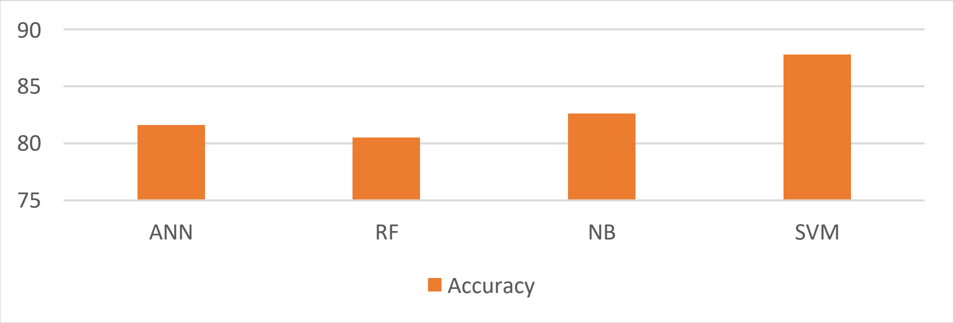

Fig. 2: sentiment classification of proposed system with various machine learning algorithms using TF-IDF Features

The above figure 2 describes classification accuracy with numerous machine learning (ML) algorithms with TF-IDF extraction of feature and selection approach. The different thresholds have been utilised during the selection of features from the entire data set, while some quality thresholds have been used for the selection of features. The data set contains numerous class labels; but due to density of entire class labels, we need to change the selection threshold accordingly. This experiment concludes that SVM provides the highest accuracy as 87.80% which is higher than other machine learning classification algorithms for both aspect identification as well as sentiment classification.

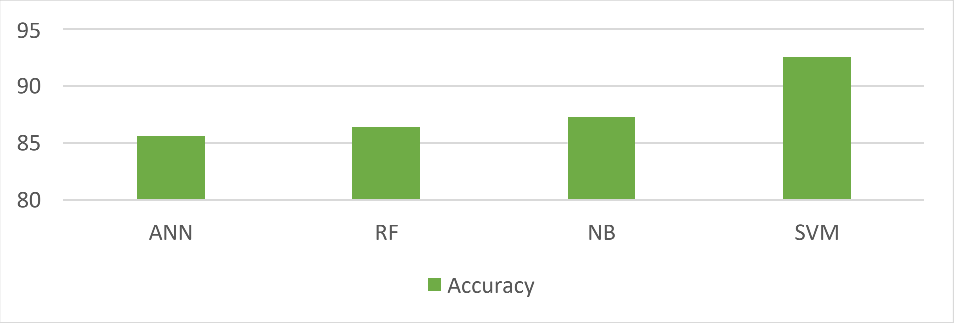

Fig. 3: sentiment classification of proposed system with various machine learning algorithms using N-Gram Features

With the N-Gram feature extraction and selection technique, the above figure 3 depicts classification accuracy with a various machine learning (ML) methods. Different criteria were used to choose features from the complete data set, and certain quality thresholds were used to select features. Because the data set includes a large number of class labels, we must adjust the selection threshold to account for this. SVM delivers the maximum accuracy of 92.50 percent, which is greater than other machine learning classification algorithms in terms of both aspect like identification and sentiment classification, according to this experiment.

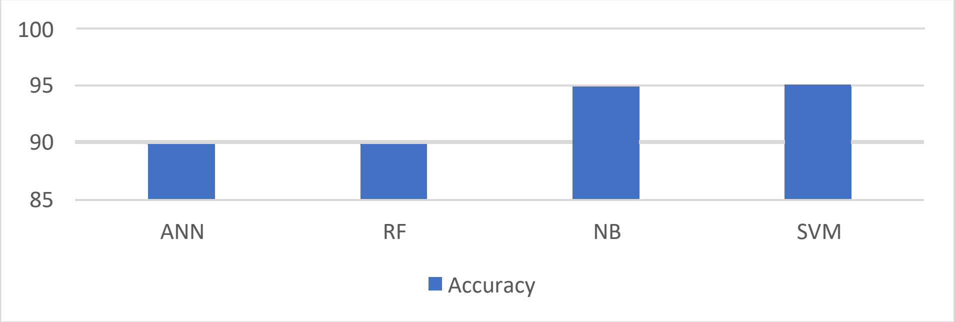

Fig. 4: sentiment classification of proposed system with various machine learning algorithms using NLP Features

There are several machine learning methods that use NLP features to extract and choose classification accuracy, as seen in the Figure 4 above. Quality criteria were employed for feature selection whereas a variety of thresholds were applied for feature selection from the complete data set. Our selection criteria must be adjusted because of the high density of all class labels in the data set. We found that SVM outperformed other machine learning classification methods in terms of aspect identification and sentiment classification, with an accuracy of 95.50 percent.

The complete experiment illustrates the distribution of the testing dataset per aspect class, with each part displaying the number of occurrences of the dataset in percentages. For cross fold data validation, 70 to 30 data split standard was applied. We gave the system roughly 5000 examples to train and 1500 instances to test the dataset.

5. Conclusion

Identifying emotions expressed by online users is beneficial to company owners, consumers, and other users, and text analytics is recognized as one of the most important subareas in text classification. Researchers noticed the requirement for our systems to capture feelings stated toward certain review components after realizing that generic emotions derived from textual data were inadequate. The system is an issue that requires supervised learning. The feature set selected and the categorization classifiers used influence systems effectiveness. The goal of this work was to increase the proposed systems overall performance. The SVM with NLP feature extraction provides 95.50% classification accuracy for both aspect identification as well as sentiment classification. The proposed research should be carried out with different feature selection algorithms on various aspect level dataset and validate the system performance on different machine learning algorithm compare with traditional approaches. It should create much more benchmark datasets from a wider range of fields in order to derive more secure findings.

References

[1] Cambria, E. Affective Computing and Sentiment Analysis. IEEE Intell. Syst. 2016, 31, 102–107.

[2] Kaur, A.; Kaur, K. Statistical Comparison of Modelling Methods for Software Maintainability Prediction. Int. J. Softw. Eng. Knowl. Eng. 2013, 23, 743–774.

[3] Al-Ghuribi, S.M.; Noah, S.A.M.; Tiun, S. Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews. IEEE Access 2020, 8, 218592–218613.

[4]. Hepburn, J. Universal Language model fine-tuning for patent classification. In Proceedings of the Australasian Language Technology Association Workshop, Dunedin, New Zealand, 11–12 December 2018; pp. 93–96.

[5] Katwe, P.; Khamparia, A.; Vittala, K.P.; Srivastava, O. A Comparative Study of Text Classification and Missing Word Prediction Using BERT and ULMFiT. In Evolutionary Computing and Mobile Sustainable Networks; Springer: Berlin/Heidelberg, Germany, 2021; pp. 493–502.

[6] Shu, K.; Bhattacharjee, A.; Alatawi, F.; Nazer, T.H.; Ding, K.; Karami, M.; Liu, H. Combating disinformation in a social media age. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2020, 10, e1385.

[7] Jakkula, V. Tutorial on support vector machine (svm). Sch. EECS Wash. State Univ. 2006, 37, 121–167.

[8] Suthaharan, S. Support vector machine. In Machine Learning Models and Algorithms for Big Data Classification; Springer: Berlin/Heidelberg, Germany, 2016; pp. 207–235.

[9] Pisner, D.A.; Schnyer, D.M. Support vector machine. In Machine Learning; Elsevier: Amsterdam, The Netherlands, 2020; pp. 101–121.

[10] Hope, T.; Resheff, Y.S.; Lieder, I. Learning Tensorflow: A Guide to Building Deep Learning Systems; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.

[11] Tarasov, D. Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews. In Proceedings of the 21st International Conference on Computational Linguistics Dialogue, Sydney, NSW, Australia, July 2015; Volume 2, pp. 53–64.

[12] Tai, K.S.; Socher, R.; Manning, C.D. Improved semantic representations from tree-structured long short-term memory networks. arXiv 2015, arXiv:1503.00075

[13] Zhang, J.; Cui, L.; Fu, Y.; Gouza, F.B. Fake news detection with deep diffusive network model. arXiv 2018, arXiv:1805.08751. 3

[14] Rani, S.; Singh, J. Sentiment analysis of Tweets using support vector machine. Int. J. Comput. Sci. Mob. Appl. 2017, 5, 83–91

[15] Saikat Bose, Tripti Arjariya, Anirban Goswami, Soumit Chowdhury Multi-Layer Digital Validation of Candidate Service Appointment with Digital Signature and Bio-Metric Authentication Approach International Journal of Computer Networks & Communications (IJCNC) Vol.14, No.5, September 2022 DOI: 10.5121/ijcnc.2022.14506