International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

# Features Extraction Effect on the Accuracy of Sentiment Classification Using Ensemble Models

**Faiza Mohammad Al-kharboush[1], Mohammed Abdullah Al-Hagery[2]**

[1, 2]Department of Computer Science, College of Computer, Qassim University, Buraidah, Saudi Arabia
[1]*381210022[at]qu.edu.sa*
[2]*hajry[atqu.edu.sa*

**Abstract:** ***A great number of works in sentiment classification have been developed, usually involving machine learning algorithms. The ensemble classifier is a subfield of machine learning that combines different base classifiers to form one powerful classifier. In the text classification, the ensemble classifier cannot process the text directly. Instead, it requires a feature extraction technique to convert the text to numeric forms. The extraction technique has great effects on the classification accuracy. The purpose of this paper is to enhance the accuracy of the ensemble classifier by defining the best feature extraction technique for the ensemble sentiment classifier. Hence, the accuracy of an ensemble model with three well-known feature extraction techniques, which are Bag of words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), Word2vec, are evaluated and analyzed on four experimental datasets. The ensemble classifier was composed of Support Vector Machine (SVM), Logistic regression (LR), k-nearest neighbor (KNN), and Random Forest (RF) as base classifiers. The analysis result indicates that using an ensemble classifier with TF-IDF delivered better classification accuracy than using BOW or word2vec. In contrast, the ensemble classifier usually reported its lowest accuracy with word2vec.***

**Keywords:** Features selection, Sentiment, Analysis, Ensemble models, classification accuracy

## 1. Introduction

Sentiment analysis has recently become a very popular and attractive research area. Sentiment analysis uses natural language processing, computational techniques, and text analysis approaches to classify text into positive and negative [1]. As a result of the vast amount of text data, the analysis of this data has become important for most governments, companies, universities, business entrepreneurs, research centers, etc... The main benefit of these analyses is to obtain feedback information about events, products, people, and services that provide helpful information for decision making [2]. The machine learning approach has gained popularity in the sentiment analysis area. An effective subfield of machine learning, i.e. ensemble learning, refers to the development of learning models by combining a diverse set of learning algorithms for increasing predictive power and performance.

In the text classifications, machine learning classifiers cannot process the raw text directly. Instead, the text must be transformed into numeric data. For this reason, feature extraction is a significate requirement for text classification. Feature extraction is an algorithm or a model that converts text to numeric forms called features. Many models have been utilized to extract features from text data such as BOW, Term Frequency–Inverse Document Frequency (TF-IDF), and Word to Vector (word2vec). Although many techniques are available, however, the suitable one must be utilized to enhance the accuracy of ensemble learning. The literature shows evidence that using suitable feature extraction technique can be enhancing the classifier performance [3]. Also, there is no agreement on the optimal feature extraction technique for all classifiers. That means if a feature extraction technique able to achieve good performance with a classifier, there is no guarantee to have the same impact with other classifiers. This paper aims to determine the best feature extraction technique for enhancing ensemble classifier performance. To this end, an ensemble classifier was evaluated with three popular techniques including BOW, TF-IDF, and word2vec for comparing their impacts. Four different datasets were used for experiment purpose. The experiments involved; gathering datasets, preprocessing, feature extraction, implicating ensemble classifier, and evaluation of the results. The rest of this paper is organized as follows: Section 2 shows the literature review. The methodology is explained in section 3. While section 4 provides the results and discussion. Finally, section 6 highlights the conclusion and future work.

## 2. Literature Review

Although the most common classifiers are working well and giving high accuracy results in different domains, for example, the applications in [4], [5], [6], [7], however, the ensemble classifiers have been proved themselves to be effective when compared with that provided as a single classifier, specifically in the sentiment classification domain. The literature shows different feature extraction techniques have been implemented with the ensemble classifier such as BOW, TF-IDF, and word2vec.

BOW is one of the well-known feature extraction techniques that is frequently used with a single machine learning classifier as well as with ensemble classifiers. The key idea of BOW is to represent the text into a collection of N numbers called vector counts. These vectors contain words and their frequency counts, meaning that BOW does not preserve the original text structure resulting from disregarding word order and grammar. Onan et al. [8] provided an ensemble scheme for sentiment classification, in which the BOW was used for feature extraction. They used five algorithms as a base classifier, which are Bayesian logistic regression, Naive Bayes (NB), linear discriminant

analysis, Logistic Regression (LR), and Support Vector Machine (SVM). Their ensemble classifier was evaluated on 9 different datasets and the results showed good achievements. Perikos et al. [9] presented an aspect-based sentiments classifier which aims to identify the sentiment of the text based on the specific aspect. The aspect-based sentiment classification task has been executed in two steps which are aspect term extraction by using a combining of various natural language processing techniques including BOW, part of speech tagging, and Stanford parser optimization and sentiment classification that combining NB, SVM, and Maximum Entropy (EM) as ensemble classifier.

TF-IDF is one of the significant techniques that is a scaled-up model of the BOW approach. It presents normalized counts of the words, in which the counts of each word are divided by the number of documents in the dataset that contain this word. The counts of words are high when the words high-frequency in the document, but decrease if the word is high-frequency in all documents in the dataset. According, the words that frequently occur in all classes is irrelevant and assigned a low number. Several previous studies confirm the preference of TF-IDF over other feature extraction techniques for some individual classifiers. For example, Wang et al [3]. examined the effects of four feature methods, i.e. word2vec, TF-IDF, doc2vec, and the counter vectorizers on SVM, LR, NB, and KNN algorithms. The TF-IDF and counter vectorizers features have maximum accuracy. The highest accuracy was achieved by SVM and LR with TF-IDF or counter vectorizer. For ensemble classifiers, many earlier works considered the TF-IDF for feature extraction [10], [11], [12].

Word2vec (word to vector) is a state of art model developed in 2013 by Google [13]. This model uses the Neural Networks approach to detect the sematic and semantic relations of the word by exploring word co-occurrence in documents given for training the model. This model represents each word as a vector located in a high dimensional vector space. If two words share a common context and have a similar meaning, then these words are placed close to each other in the vector space. Al-azani[14] recommended the use of the word2vec feature with ensemble classifiers for sentiment classification in the case of imbalanced class datasets. Also, Xu [15] utilized the word2vec with an ensemble classifier, which combined Neural Networks models, for sentiment classification. The evaluation results confirm the effectiveness of this framework, in which it achieved a substantial performance gain more than its base models. Since the ensemble classifiers were utilized with different feature extraction techniques in the literature, this paper investigates the best one for enhancing the classification performance.

## 3. Methodology

This section explains the overall steps of the methodology used in this paper, which involve gathering and preprocessing the datasets, feature extraction, ensemble classification, and evaluations. Fig. 1 illustrates these steps.

### 3.1 Preprocessing

Preprocess steps are the processes of removing redundant and irrelevant information to decrease the feature size. These steps allow enhancing the accuracy of machine learning algorithms [20].

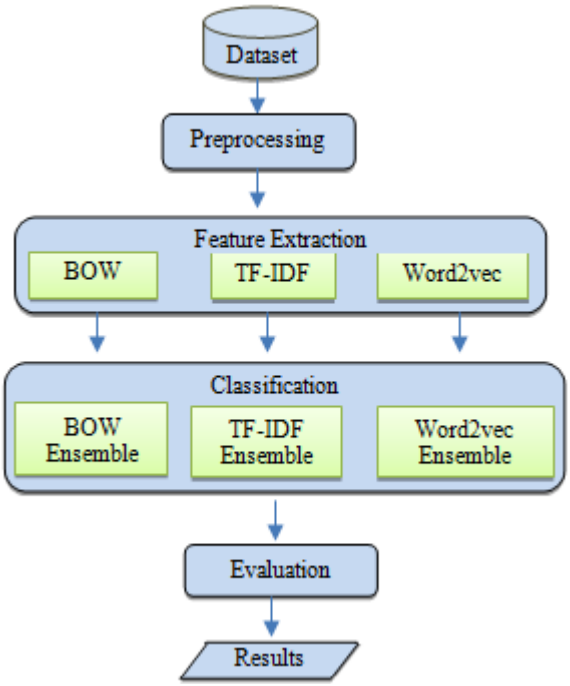


**Figure 1:** Methodology steps

### 3.2 Datasets Gathering

The collected datasets involve different sizes and sources including Twitter datasets, news datasets, and reviews datasets. Table 1 shows a brief description of each dataset.

**Table 1:** Dataset description

| # | Dataset name | size | Positive | Negative | Source |
|---|---|---|---|---|---|
| 1 | health care reform (hcr)[16] | 1,066 | 729 | 337 | Twitter data |
| 2 | Sander [17] | 832 | 317 | 515 | Twitter data |
| 3 | Financial Phrase Bank-100 (FP)[18] | 873 | 507 | 303 | News data |
| 4 | Book [19] | 2000 | 1000 | 1000 | Review data |

Five popular preprocessing steps of text data, i.e. data cleaning, stop word removal, lowercase conversation, tokenization, and stemming are considered in this work. A brief definition of these steps as follows:

- Data cleaning: To improve the quality of data, unwanted observations such as punctuations, numbers, short words, and special characters, were removed.
- Stop words removal: Stop words are high-frequency words without dependency on a specific topic such as prepositions. Stop words are commonly assumed to be redundant and irrelevant in sentiment analysis studies because they have high occurrence rates without giving useful meaning.
- Lowercase conversation: Since lowercase and uppercase forms of words are expected to have no different meaning, all capital letters were converted to small letters.

- Tokenization: is a process of segmenting the text into meaningful parts such as words or phrases, namely tokens.
- Stemming: is a process is to get the root forms of all words. Stemming is important in natural language processing to deal with the different forms of the derived words as a single stem word.

### 3.3 Feature Extraction

Among different feature extraction techniques, the BOW is the most common domination technique utilized with ensemble learning in previous studies. However, some of these studies have confirmed that the TF-IDF technique achieves better results than BOW [3]. The BOW and TF-IDF focus on the word frequency and treat the words as discrete symbols. Thereby, they don't take into consideration the word order and semantic of the words. The word2vec overcome this limitation by capturing the additional information of the word such as word similarity. Hence, BOW, TF-IDF, and word2vec are selected as experiment variables. By using Python programming language, the BOW and TF-IDF techniques were implemented through utilizing sci-kit-learn library, while the gensim library was used to import word2vec model. Each resulted features were divided into training and testing sets as 80% and 20%, respectively. The training set was served training the ensemble classifier and the testing set was used for evaluation purposes.

### 3.4 Ensemble Classification and Evaluation

To meet the main goal of these experiments, an ensemble classifier is constructed and evaluated with all constructed features. For building an ensemble model, we have to define the base classifiers and the combination method of these classifiers. Among different combination methods, voting is an effective, simple, and frequently used method for forming ensemble learning. In the simplest voting approach, named majority voting, each one of the base classifiers set is contributed equally and provide a single vote.

The most frequented vote is considered as the final classification output of the ensemble model. So, the majority voting method was used as a combination method. The SVM, Logistic regression (LR), k-nearest neighbor (KNN), and Random Forest (RF) were included as base classifiers for the ensemble model because they are popular and well-known base classifiers in sentiment classification. The ensemble classifier was trained using the training subset of each constructed feature. For evaluation purposes, the accuracy metric was considered as an evaluation measure on testing subsets of each feature. The accuracy is a ratio of true classification instance among all instances.

## 4. Result and Discussion

Based on the experiments, we evaluated the ensemble model and its base classifier accuracy with BOW, TF-IDF, and wor2vec features as shown in Table 2.

**Table 2:** Classifiers evaluation among different features

| Dataset | Feature | Classifier Name | | | | Ensemble Model Accuracy |
|---|---|---|---|---|---|---|
| | | LR | SVM | KNN | RF | |
| Hcr | BOW | 78.29 | 76.35 | 74.41 | 78.68 | 77.51 |
| | TF-IDF | 75.96 | 76.74 | 74.41 | 75.58 | 77.51 |
| | Word2vec | 79.06 | 76.74 | 76.35 | 77.13 | 77.13 |
| Sander | BOW | 80.83 | 70.65 | 76.64 | 61.67 | 79.64 |
| | TF-IDF | 80.83 | 76.04 | 75.44 | 75.44 | 80.23 |
| | Word2vec | 75.44 | 67.66 | 69.46 | 63.47 | 69.46 |
| FP | BOW | 78.85 | 84.00 | 79.42 | 73.71 | 79.42 |
| | TF-IDF | 77.71 | 80.57 | 77.71 | 77.14 | 80.00 |
| | Word2vec | 75.42 | 62.85 | 72.00 | 64.57 | 73.71 |
| Book | BOW | 74.25 | 70.50 | 75.25 | 55.75 | 72.5 |
| | TF-IDF | 77.00 | 66.25 | 78.25 | 69.75 | 77.00 |
| | Word2vec | 73.00 | 69.00 | 74.25 | 67.50 | 74.50 |

In Table 2, the best feature for the ensemble model and its base classifiers on all used datasets are highlighted. The findings confirm the usefulness of the TF-IDF feature over BOW and Word2vec features for the ensemble model in all datasets. On the other hand, the Word2vec feature is often the worst choice for the ensemble model in all datasets except for the Book dataset. For base classifiers, it is clear that no single feature can uniformly outperform other features over all datasets. For example, there is no agreement on the best feature in the case of SVM according to the tested datasets. Further analysis showed that the best feature differs based on the used classifiers and dataset. However, TF-IDF confirms its effectiveness in most circumstances of various base classifiers and datasets.

Many interesting observations highlight the relationship between the ensemble model and its base classifiers in terms of feature influences. Firstly, we can see in the case of the Hcr dataset that no one of the base classifiers was preferred with TF-IDF, however, their ensemble model was recorded its best result with the TF-IDF feature. Secondly, most base classifiers in Financial Phrase Bank-100 dataset reported their best performance with BOW, while their ensemble performed its greatest performance with TF-IDF.

Accordingly, these observations indicate that the TF-IDF feature is suitable for ensemble learning regardless of the optimal feature for its base classifiers.

The usefulness of the TF-IDF over others maybe because it highlights significant words more than the BOW and Word2vec features. This means TF-IDF is distinct from BOW in giving the frequent words in all datasets, positive and negative data, lower values than the distinctive words. Thus, TF-IDF accentuates the informative words and eliminates the common word [21]. Under the assumption that the Word2vec represents extra semantic features, Word2vec neglects the frequency of each word concerning the used dataset. The distinction of unique words for each class is helpful in text classification. This is just a supposed reason and further work is needed to explore the theoretical reasons behind this result.

## 5. Conclusion and Future Work

Since any machine learning classifier requires an extraction technique that converts the text into a convenient form called features. This work examines three well-known

extraction techniques, which are BOW, TF-IDF, and word2vec, to define the convenient one for ensemble learning. Different scenarios of applying the three extraction techniques with the ensemble model were evaluated on four experimental datasets. The experiment results reveal that the TF-IDF technique yields enhancement on ensemble accuracy more than others. On other hand, the word2vec technique is usually the worst choice for ensemble learning. As future work, the experiments can be expanding by testing more than three feature extraction methods with more than four testing datasets. Also, we can examine more than one type of ensemble classifiers by changing the combination method. Besides, by using the association rules methods [22], it can find a set of similar and associated classifiers that can give high results as an ensemble model based on using the Apriori algorithm.

## References

[1] D. Mohey and E. M. Hussein, "ORIGINAL ARTICLE A survey on sentiment analysis challenges," *J. King Saud Univ. - Eng. Sci.*, vol. 30, no. 4, pp. 330–338, 2018.
[2] Y. Yaslan and D. Aldo, "A comparison study on active learning integrated ensemble approaches in sentiment analysis R," vol. 57, pp. 311–323, 2017.
[3] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu, "Comparisons and Selections of Features and Classifiers for Short Text Classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 261, no. 1, 2017.
[4] M. A. H. Al-Hagery, "Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques," *Int. J. Adv. Biotechnol. Res.*, vol. 7, no. 2, pp. 760–772, 2016.
[5] S. Al-qarzaie, S. Al-odhaibi, B. Al-saeed, and M. Al-hagery, "Using the Data Mining Techniques for Breast Cancer Early Prediction," *Symp. Data Min. Appl.*, vol. 1, no. May, 2014.
[6] A. Abdulrahman Al-Noshan, M. Abdullah Al-Hagery, H. Abdulaziz Al-Hodathi, and M. Sulaiman Al-Quraishi, "Performance Evaluation and Comparison of Classification Algorithms for Students at Qassim University," *Int. J. Sci. Res.*, vol. 8, no. 11, pp. 1277–1282, 2018.
[7] E. I. Al-Fairouz and M. A. Al-Hagery, "The most efficient classifiers for the students' academic dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 501–506, 2020.
[8] S. Koruko, "A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification," vol. 62, pp. 1–16, 2016.
[9] I. Perikos, "Aspect based Sentiment Analysis in Social Media with Classifier Ensembles," *2017 IEEE/ACIS 16th Int. Conf. Comput. Inf. Sci.*, pp. 273–278, 2017.
[10] H. M. Abdelaal, A. N. Elmahdy, A. A. Halawa, and H. A. Youness, "Improve the automatic classification accuracy for Arabic tweets using ensemble methods," *J. Electr. Syst. Inf. Technol.*, no. 2017, pp. 1–8, 2018.
[11] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classi fi cation : The contribution of ensemble learning," vol. 57, pp. 77–93, 2014.
[12] M. AbdullahAl-Hagery, M. AbdullahAl-Assaf, and F. MohammadAl-Kharboush, "Exploration of the best performance method of emotions classification for arabic tweets," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 19, no. 2, pp. 1010–1020, 2020.
[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Prepr. arXiv1301.3781*, 2013.
[14] S. Al-azani, "Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Sentiment Analysis in Short Arabic Text,," *Procedia Comput. Sci.*, vol. 109, pp. 359–366, 2017.
[15] X. Xu, "UNIMELB at SemEval-2016 Tasks 4A and 4B : An Ensemble of Neural Networks and a Word2Vec Based Model for Sentiment Classification," pp. 183–189, 2016.
[16] M. Speriosu, S. Upadhyay, and J. Baldridge, "Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph," pp. 53–63, 2011.
[17] N. J. Sanders, "Sanders-twitter sentiment corpus," *Sanders Anal. LLC*, vol. 242, 2011.
[18] P. Malo, A. Sinha, P. Takala, P. Korhonen, and J. Wallenius, "FinancialPhraseBank-v1.0." 2013.
[19] J. Blitzer, "Biographies , Bollywood , Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification," 2006.
[20] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Comput. Math. Organ. Theory*, no. 2, pp. 1–17, 2018.
[21] J. Lilleberg, Y. Zhu, and Y. Zhang, "Support vector machines and Word2vec for text classification with semantic features," *Proc. 2015 IEEE 14th Int. Conf. Cogn. Informatics Cogn. Comput. ICCI*CC 2015*, pp. 136–140, 2015.
[22] M. A. Al-Hagery, "Extracting hidden patterns from dates' product data using a machine learning technique," *IAES Int. J. Artif. Intell.*, vol. 8, no. 3, pp. 205–214, 2019.

## Author Profile

**Faiza Mohammad Al-Khrboush** received the B.Sc. degree in Computer Science from the University of Technology, Qassim University, Saudi Arabia, in 2016. Currently, she is aComputer Science Master student at Qassim University. She published one research paper in 2020.

**Dr. Mohammed Abdullah Al-Hagery** received the B.Sc. degree in computer science from the University of Technology, Baghdad, Iraq, in 1994, the M.Sc. degree in computer science from the University of Science and Technology (USTY), Sana'a, Yemen, in 1998, and the Ph.D. degree in Computer Science and Information Technology (Software Engineering) from the Faculty of Computer Science and IT, University Putra Malaysia (UPM), in November 2004. He was the Head of the Computer Science Department, College of Science and Engineering, USTY, from 2004 to 2007. Since 2007, AL-HAGERY has been a Staff Member in the Department of Computer Science, College of Computer, Qassim University, Saudi Arabia. He was appointed as the Head of the Research Centre, Computer College, and a Council Member of the Scientific Research Deanship, Qassim University, from September 2012 to October 2018. He is currently an Associate Professor, has published more than 33 research papere in various international journals. AL-HAGERY is teaching the master's degree students and a supervisor of many master's dissertations. He is a Jury Member of several Ph.D. and master's thesis as well he is an internal and external examiner in the field of his specialization