DEU FMD 26(78) (2024) 440-448

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

A Text Mining Application Using Weighted Majority Voting Ensemble Method

Ağırlıklı Çoğunluk Oylama Topluluğu Yöntemini Kullanan Bir Metin Madenciliği Uygulaması

Alican Doğan [1*], Mansur Alp Toçoğlu [2]

[1] Bandırma Onyedi Eylül Üniversitesi Uygulamalı Bilimler Fakültesi Yönetim Bilişim Sistemleri Bölümü, Balıkesir, TÜRKİYE
[2] İzmir Katip Çelebi Üniversitesi Mühendislik ve Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü, İzmir, TÜRKİYE

*Sorumlu Yazar / Corresponding Author **: alicandogan@bandirma.edu.tr

Abstract

In text mining, sentiment analysis is gaining popularity day by day although it has been recently introduced. One of the important feedback parameters of this research is the opinion about text-based content. The general goal in this aspect is to analyze product and service reviews or comments so that they can be compared and contrasted with each other via the ratings they get. An ensemble method which we have proposed earlier is used in this study to boost the classification accuracy of different conventional single machine learning models. Five analytical models that are related but not identical are implemented and their class decisions are integrated using a special weighted majority voting ensemble mechanism called WMVE to increase the classification score of the data mining technique. Naïve Bayes, OneR, Hoefding Tree, REPTree, and KNN methods are utilized as base classifiers in the ensemble and their class decision are integrated into the WMVE method. At the same time, outputs were compared to the ones obtained by Standard Majority Voting Ensemble (MV) including the same base classifiers. Based on the findings, the WMVE model demonstrated superior performance compared to other classifiers, achieving an average accuracy of 77.35 and F-Score of 77.19 values. Consequently, the ensemble model including WMVE is used to enhance sentiment analysis classification performance.

Keywords: *Majority voting, text mining, classification, ensemble*

Öz

Metin madenciliğinde, yakın zamanda tanıtılmasına rağmen duygu analizi gün geçtikçe popülerlik kazanmaktadır. Bu araştırmanın önemli geri bildirim parametrelerinden biri, metin tabanlı bir içerik hakkındaki görüşlerdir. Bu konudaki genel amaç, ürün ve hizmet incelemelerini veya yorumlarını, aldıkları puanlar aracılığıyla birbirleriyle karşılaştırabilmeleri ve karşılaştırabilmeleri için analiz etmektir. Farklı geleneksel tek makine öğrenimi modellerinin sınıflandırma doğruluğunu artırmak için bu çalışmada daha önce önerdiğimiz bir topluluk yöntemi kullanılmıştır. Veri madenciliği tekniğinin sınıflandırma puanını artırmak için birbiriyle ilişkili ancak aynı olmayan beş analitik model uygulanmış ve bunların sınıf kararları, Ağırlıklı Çoğunluk Oylaması (WMVE) adı verilen özel ağırlıklı çoğunluk oylama topluluğu mekanizması kullanılarak entegre edilmiştir. Toplulukta temel sınıflandırıcılar olarak Naïve Bayes, OneR, Hoefding Tree, REPTree ve KNN yöntemleri kullanılmış ve bunların sınıf kararı WMVE yöntemi için entegre edilmiştir. Aynı zamanda sonuçlar aynı sınıflandırıcılarla oluşturulan Standart Çoğunluk Oylaması (MV) bulgularıyla da kıyaslanmıştır. Bulgulara göre, WMVE modeli, diğer sınıflandırıcılara kıyasla üstün performans sergiledi ve ortalama doğruluk değeri olarak 77.35 ve F-Skoru olarak 77.19 değerlerine ulaştı. Sonuç olarak, duygu analizi sınıflandırma doğruluğunu artırmak için ağırlıklı oylama yöntemini içeren topluluk modeli kullanılır.

Anahtar Kelimeler: *Çoğunluk oylaması, metin madenciliği, sınıflandırma, topluluk*

1. Introduction

The content created by users through information input is one of the most preferred accessible data used in the evaluation of a product or service. These contents are formed by users expressing their experiences about products or services in digital media. With the widespread use of web technologies, manufacturers, business owners, and sellers refer to their users' feedback data to gain a commercial advantage [1].

With the growth of online commerce and social networking sites, a significant amount of product or service evaluation data has been collected. In the referred community-based approach, the proposed method integrates supervised machine learning techniques with majority voting. The objective of this approach is to optimize the decisions made by many classifiers to improve the classification quality. Voting is conducted among five classifiers: Naïve Bayes, ZeroR, BayesNet, Multinomial Text, and Logistic Regression. In these experiments, the majority voting community-based approach outperforms all individual classifiers and the standard majority voting [2].

Online opinions are generated for different purposes and have different side effects [3]. For example, a positive opinion can lead to financial growth, while a negative opinion can cause a decrease in sales. Therefore, sellers welcome the use of user opinions to improve their business. They analyze this feedback and adjust their products and services accordingly. Sentiment analysis constitutes a significant area of investigation in text mining,

DOI: https://doi.org/10.21205/deufmd.2024267810

Geliş Tarihi / Received: 20.10.2023
Kabul Tarihi / Accepted: 06.01.2024

natural language processing, and information retrieval. It serves the purpose of succinctly representing text documents, offering advantages in various applications such as automatic indexing, summarization, classification, clustering, and filtering. One specific application is in the domain of text classification, where the challenge of a high-dimensional feature space can be addressed by extracting the most crucial or relevant words from the document content and utilizing them as features.

Sentiment analysis is implemented in three levels. These levels are archive, sentence, and viewpoint level. In our study, we take care of sentences. Record-level organization determines whether the document's viewpoint is favorable, unfavorable, or impartial. The sentence level determines whether the sentence conveys a regrettable, positive, or neutral evaluation. The field of text classification presents a challenge due to the high-dimensional feature space [4].

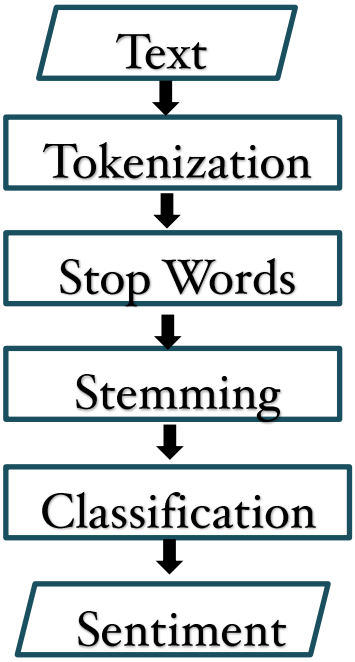

Figure 1. Sentiment analysis process.

Perspective level examines the entirety of the emotions expressed within the provided document and the perspective it refers to. Sentiment analysis process is demonstrated above in Figure 1.

The challenge of high-dimensional feature space is a common issue in text classification applications. Utilizing all words from training documents as features makes the text classification process computationally intensive. Therefore, selecting keywords from a text collection, representing the most crucial and relevant words in the document content, becomes a favorable choice for constructing features in a classification model. Machine learning algorithms like Naïve Bayes, k-nearest neighbor, support vector machines, and artificial neural networks have proven successful in text document classification. Ensemble methods, a collection of learning algorithms, combine decisions from these algorithms to build a more robust classification model with enhanced predictive performance.

2. Related Work

In the literature, there are studies focusing on analyzing the customer reviews obtained from Amazon in terms of categorizing the sentiments of them. Within the scope of this section, we briefly explain several studies related to sentiment analysis of Amazon customer reviews. Text format eases the process of recording information but poses difficulties when attempting to utilize the data for secondary purposes [5].

In the study [6], the authors focused on the evaluation of sentiment analysis of balanced and unbalanced Amazon online review datasets. To do so, they used four well-known deep learning algorithms which are Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Group Long Short-Term Memory (GLSTM), Gated Recurrent Unit (GRU) and Updated Recurrent Unit (URU). They utilized 3 different word embedding feature extraction methods which are Glove, Word2vec and FastText. The macro average prediction results were evaluated based on accuracy, recall, precision and F1-score. They achieved the highest accuracy result as 93.73 for the unbalanced dataset using GLSTM network with FastText feature extraction method. On the other hand, the LSTM networks provided the highest accuracy performance as 88.39 for the balanced dataset.

The authors analyzed sentiments of Amazon customer reviews in their study [7]. They used the Amazon Review dataset 2018 which consists of 938254 records from the accessories and cell phone section of Amazon in total. They compared the classification performances of two conventional machine learning classifiers (Naive Bayes and Support Classification Algorithm) with two deep learning classifiers (LSTM and Convolutional Neural Network(CNN)) in terms of accuracy, precision, recall and F1-score. According to the results, LSTM provided the highest performance among other classifiers with 93% accuracy and 97% F1-score values.

In another study [8], the sentiment polarization of the Amazon product reviews had been analyzed by using two conventional machine learning classifiers which are Naive Bayes (NB) and Support Vector Machine (SVM) in terms of four metrics (accuracy, recall, precision and F1-score). The authors used a raw dataset which is composed of more than one million Amazon product feedbacks. The dataset is automatically labeled according to 5-star rating system. Considering the classification results, SVM outperformed NB in terms of all metrics.

Khalid et al. [9] proposed a voting classifier named Gradient Boosted Support Vector Machine (GBSVM) based on two base models which are Gradient Boosting and Support Vector Machine. They utilized a dataset containing 64,295 records which are the mobile application reviews for Google apps. In the experimental section of the study, the authors focused on investigating prediction performances of various machine learning classifiers (Support Vector Machine, Gradient Boosting Machine, Logistic Regression and Random Forest), the proposed voting classifier and four state-of-the-art ensemble methods. The outcome of the results indicated that the proposed model outperformed other classifiers.

Qorich and Ouazzani [10] focused on extracting positive and negative sentiments of Amazon customer reviews within the scope of their study. To do so, they proposed a CNN model and compared its performance with baseline machine learning and deep learning classifiers in terms of accuracy, precision, recall and F1-score. They also implemented experiments by using diverse model designs. The overall results indicated that the proposed CNN model achieved higher performance as 90% accuracy value among other classifiers.

In the study [11], the authors performed an aspect level sentiment analysis considering bipolar words on Amazon product reviews. To do so, they first used Scrapy to collect the raw dataset which is composed of 191,720 reviews of six different products and then they performed four stages to

preprocess the data. Next, the authors identified the bipolar words with their adjustment values using aspect level sentiment analysis. After the identification of bipolar words, the classification phase is performed by using the support vector machine classifier with its three different kernels which are linear, polynomial and radial basis function (RBF). According to the classification performances, RBF provided the best results among others.

Alroobaea [12] studied sentiment analysis on Arabic Amazon product reviews using deep learning architectures. The author used three different datasets with different sizes. After the pre-processing phase, a recurrent neural network (RNN) model was applied to predict reviews in terms of positive and negative sentiments. The results obtained by the proposed RNN model were compared to three well-known deep learning architectures which are Long short-term memory (LSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN). According to the overall prediction results, the proposed model achieved the highest accuracy values in all three datasets.

3. Materials and Method

The main objective of the research is to propose a novel ensemble method with a special weighted majority voting mechanism to improve the predictive performances achieved in a sentiment analysis problem. The overall view of the proposed method is revealed in Figure 2.

As shown in Figure 2 [13], the first stage of the proposed framework is selecting single classification algorithms. In our study, we have chosen 5 different classifiers. This number can be increased or decreased according to the problem. The algorithm assigns distinct weights to each trained classifier based on how well they perform on the validation set. The ultimate prediction for each instance is determined by considering the votes of the classifiers with the highest weights [13].

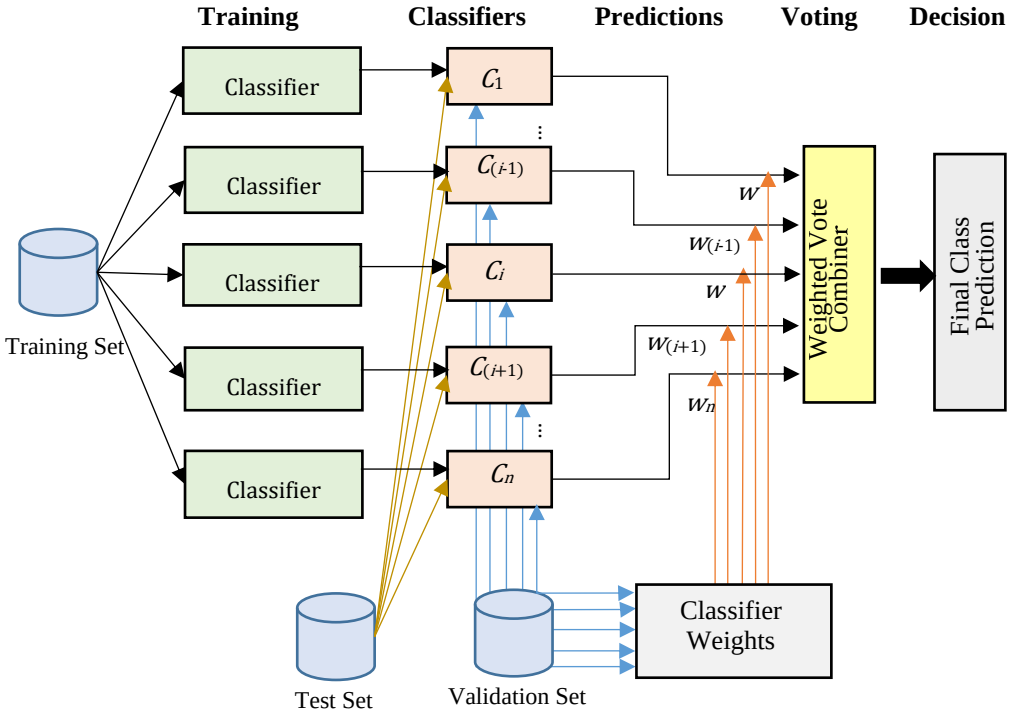

Figure 2. The structure of the proposed method

The next phase of the proposed method is to determine the weights of each classifier. In this step, all objects in the validation dataset are traversed and processed through n classifiers once. Then, the weights of the classification methods having correct predictions are increased by the ratio of the number of incorrectly predicting classifiers to the whole number of classifiers (n). The output of this step is the classifiers with updated weights to vote decision of class label of each instance in the test test.

Different text mining preprocessing steps such as normalization, tf-idf transformation, bag-of-words, removing stop words etc. are implemented on the training set. This weight gain value cannot exceed 1 for a single instance operation. Those preprocessing operations have been performed using python machine learning libraries and Weka [16] filtering tools. Following the implementation of these preprocessing procedures, the result is subsequently directed to the subsequent phase, known as the classification stage.

In the WMVE approach that is being suggested, there consist of three distinct stages. The initial stage involves training classifiers using a training set. Subsequently, the second stage involves ascertaining the weights of these classifiers by utilizing a validation set. During this phase, each classifier produces a decision pertaining to the predicted class label of a single instance, and these decisions are assessed to adjust the weights. Lastly, in the third stage, the outputs from individual classifiers are amalgamated, taking into account their respective weights.

Therefore, the complete dataset is divided into three distinct parts: the training set, the validation set, and the test set. The validation set's size matches that of the test set, amounting to one-eleventh of the entire dataset. Let's denote "m" as the number of instances within the validation set. Table 1 provides a visual representation of the classifiers' accurate and inaccurate predictions for each instance in the validation set, where

$$p_{ij} = \begin{cases} 1, & \text{if } j^{th} \text{ classifier makes a correct prediction for } i^{th} \text{ instance} \\ 0, & \text{if } j^{th} \text{ classifier makes an incorrect prediction for } i^{th} \text{ instance} \end{cases}$$

Table 1. Predictions of classifiers in the validation set.

Instances	C1	C2	...	C_n
1	p_{11}	p_{12}	...	p_{1n}
2	p_{21}	p_{22}	...	p_{2n}
.	.	.	.	.
m	p_{m1}	p_{m2}	...	p_{mn}

In the classification stage, five different single classifiers, normal majority voting ensemble, and the proposed weighted majority voting ensemble methods consisting of these five classifiers are implemented. These classifiers are OneR, Hoefding Tree, REPTree, k-NN, and Naive Bayes methods. All models were applied using C# code snippets and related Weka classification libraries.

4. Experimental Results

In this study, we used a pre-collected dataset from Kaggle which is composed of 10 columns and 568,454 rows in total. The dataset contains customer reviews of fine foods from Amazon [14]. We made several changes on the dataset to make it ready for the tasks implemented in this paper. Firstly, we separated the review and the score columns and created a sub-dataset. Next, we redefined the rating values as negative and positive within the score column where the value range is between 1 to 5 indicating 1 as the lowest rating value and 5 as the highest rating value. We set 1 and 2 rating values as negative (0) and 5 as positive (1). After the labeling process, the sentiment distribution of the dataset became 363,122 positive and 82,037 negative reviews. However, we shrunk the dataset to 4,938 positive and 4,703 negative reviews due to the calculation problems of the proposed methods in this study. Table 2 shows several customer review examples with their label.

After the re-construction of the dataset, we applied pre-processing methods on the dataset to make it ready to be used in experimental procedures. First, we converted all characters to lowercase. Second, we removed punctuation characters, the extra spaces, and all numeric characters. Then, we normalized each term in the dataset by using SnowballStemmer. At last, we removed all stopwords defined within the version 3.7 NLTK library [15].

In this phase of the study, we compared the macro average predictive performances of the proposed model WMVE with six different machine learning classifiers (OneR, HT, REPTree, NB, KNN and MV) in conjunction with four different stemming approaches (namely, Snowball, IL, Lovins and None). We achieved all experimental results by using TF-IDF text representation scheme. Five different ngram models have been evaluated ranging from 1 to 3. In addition, we considered the combination of ngram models which are unigram-bigram(uni-bi), bigram-trigram(bi-tri), and unigram-bigram-trigram (uni-bi-tri). In the empirical analysis, we set the feature size as 1,000 and utilized 10-fold cross validation where the dataset is divided into three parts in each fold which are training(9/11), testing(1/11) and validation(1/11).

We implemented a C# program to perform all the empirical analysis. We have utilized Weka machine learning library suitable for visual studio environment to figure out classification tasks. We have used default input parameters for each classifier. The application enable us to compare and contrast classification performance values obtained by single classifiers and ensemble models.

Tables 3,4,5,6 show the accuracy, precision, recall and F-score values achieved for all cases in the empirical analysis. Among all the compared configurations in terms of all four metrics, similar combinations provided the highest and the lowest scores. The proposed model WMVE using the None stemming approach in conjunction with the uni-bi-tri n-gram combination have achieved the highest predictive performances for all metrics. In contrast, the lowest predictive performances have been achieved by the classifier OneR using the Snowball stemming approach in conjunction with the bigram and bi-trim n-gram combinations.

Table 2. Customer review samples.

Customer Review	Label
I love these chips and they are so much healthier than regular chips and they taste great and they look unique	1
found this tea while living in seattle a few years ago absolutely my favorite	1
people who spend this kind of money on belgian chocolate will likely never be able to afford a vacation in belgium	0
unfortunately this is a very poor representation of jack links beef jerky the bags are small and the jerky is in very small pieces and crumbs it looks like scraps or left overs no piece is larger than inch	0
this is great i have bought it for years and it is always good nothing better with a morning cup of coffee	1
this coffee tastes like any other i highly doubt it is even jamacain blue mountain since it doesnt have that distinct taste save your money	0
my son is a chef in the making and this will be a great addition to his ingredients list he will enjoy this	1

Table 3. Accuracy values of seven classifiers for all cases.

	OneR	HT	REPTree	NB	KNN	MV	WMVE
Snowball_unigram	60.41	60.07	68.67	55.49	54.18	63.96	64.83
Snowball_(uni-bi)	60.43	70.36	68.1	54.7	53.39	69.15	69.09
Snowball_(uni-bi-tri)	60.62	69.81	68.91	54.61	53.9	69.79	69.77
Snowball_bigram	50.28	64	67.54	54.08	49.87	64.21	68.03
Snowball_(bi-tri)	50.23	68.17	68.87	52.79	51.77	64.34	68.23
IL_unigram	59.67	66.64	76.59	73.2	62.18	76.16	80.26
IL_(uni-bi)	59.19	65.89	75.97	73.95	61.89	77.46	81.95
IL_(uni-bi-tri)	59.92	64.42	75.93	74.27	61.7	77.22	80.99
IL_bigram	59.85	64.96	75.25	74.38	61.04	76.81	81.31
IL_(bi-tri)	59.57	64.25	75.38	73.93	61.35	77.22	81.16
Lovins_unigram	61.76	64.79	78.62	70.05	64.55	72.27	79.27
Lovins_(uni-bi)	62.98	63.48	77.12	70.79	64.78	72.86	79.44
Lovins_(uni-bi-tri)	60.64	64.74	78.03	71.35	63.87	72.19	77.38
Lovins_bigram	60.39	62.78	78.08	72.63	62.88	72.95	78.99
Lovins_(bi-tri)	60.9	62.67	76.48	71.05	64.25	73.54	76.79
None_unigram	68.19	69.47	77.14	73.98	59.86	75.59	80.99
None _(uni-bi)	66.77	69.07	74.66	71.45	62.62	74.19	80.25
None _(uni-bi-tri)	66.57	70.45	75.92	75.16	62.26	76.94	83.68
None _bigram	64.57	71.51	77.89	75.45	60.13	73.21	82.57
None _(bi-tri)	64.22	67.9	73.2	73.9	58.04	75.82	82.04

Table 4. Precision values of seven classifiers for all cases.

	OneR	HT	REPTree	NB	KNN	MV	WMVE
Snowball_unigram	60.88	58.79	70.1	55.66	55.97	64.59	63.77
Snowball_(uni-bi)	62.2	72.21	69.36	52.74	51.59	67.99	68.11
Snowball_(uni-bi-tri)	60.06	68.37	70.69	53.76	54.06	71.05	70.41
Snowball_bigram	48.89	62.23	66.07	55.04	51.53	63.18	67.71
Snowball_(bi-tri)	49.86	68.04	69	54.69	53.25	63.99	66.99
IL_unigram	58.5	66.29	75.65	74.38	62.2	77.19	78.63
IL_(uni-bi)	60.19	67.64	77.95	74.48	61.26	76.27	80.98
IL_(uni-bi-tri)	60	66.33	76.68	76.23	63.55	75.49	79.32
IL_bigram	61.28	66.74	74.74	76.35	61.01	76.2	80.24
IL_(bi-tri)	57.76	66.04	75.53	72.11	62.12	77.48	79.41
Lovins_unigram	62.26	65.57	77.46	69.71	65.17	72.59	80.19
Lovins_(uni-bi)	63.02	61.48	76.98	70.46	64.08	71.87	78.75

Lovins_(uni-bi-tri)	62.06	65.27	76.24	72.42	63.56	73.37	79.06
Lovins_bigram	61.93	60.99	76.91	73.1	63.56	73.37	79.62
Lovins_(bi-tri)	59.32	63.93	76.82	70.47	62.33	72.01	75.58
None_unigram	69.38	68.14	76.83	75.26	61.05	75.01	80.8
None _(uni-bi)	66.98	69.52	73.29	71.79	60.92	73.73	81.68
None _(uni-bi-tri)	68.48	71.01	75.99	75.28	61.06	75.27	84.11
None _bigram	65.06	73.03	79.88	73.58	58.15	74.17	82.59
None _(bi-tri)	64.13	68.7	71.79	74.84	56.34	75.01	83.85

Table 5. Recall values of seven classifiers for all cases.

	OneR	HT	REPTree	NB	KNN	MV	WMVE
Snowball_unigram	60.03	58.62	70.42	56.41	55.23	65.97	62.98
Snowball_(uni-bi)	62.57	71.2	68.01	52.46	51.8	67.07	67.53
Snowball_(uni-bi-tri)	60.79	68.88	69.72	54.62	54.98	72.42	70.96
Snowball_bigram	47.95	62.46	65.51	55.19	52.36	64.45	68.79
Snowball_(bi-tri)	50.87	69.19	68.14	55.52	53.77	63.62	66.98
IL_unigram	58.75	65.1	75	73.34	61.07	76.04	79.99
IL_(uni-bi)	61.33	67.04	77.34	75.91	62.73	77.74	79.92
IL_(uni-bi-tri)	60.24	65.62	77.53	76.97	62.08	75.99	79.9
IL_bigram	61.83	67.65	74.7	74.86	60.14	75.59	81.13
IL_(bi-tri)	56.87	65.99	74.33	70.94	62.22	77.03	80.76
Lovins_unigram	62.95	64.12	78.82	69.81	66.4	71.46	79.57
Lovins_(uni-bi)	62.87	62.37	76.26	69.1	63.33	70.98	77.94
Lovins_(uni-bi-tri)	60.66	64.18	76.35	73.25	64.62	72.58	78.7
Lovins_bigram	61.55	61.64	77.62	73.94	64.75	74.51	78.65
Lovins_(bi-tri)	60.07	64.66	77.3	70.48	60.95	71.59	76.16
None_unigram	68.56	69.3	77.29	75.2	62.09	73.67	81.51
None _(uni-bi)	67.79	70.99	72.28	71.96	61.7	72.68	83.15
None _(uni-bi-tri)	67.53	71.81	76.17	74.29	61.39	74.82	84.96
None _bigram	65.44	71.8	78.93	74.98	59.09	73.15	83.08
None _(bi-tri)	63.44	70.13	72.94	76.18	56.64	76.03	83.39

Table 6. F-score values of seven classifiers for all cases.

	OneR	HT	REPTree	NB	KNN	MV	WMVE
Snowball_unigram	60.45	58.70	70.26	56.03	55.60	65.27	63.37
Snowball_(uni-bi)	62.38	71.70	68.68	52.60	51.69	67.53	67.82
Snowball_(uni-bi-tri)	60.42	68.62	70.20	54.19	54.52	71.73	70.68

Snowball_bigram	48.42	62.34	65.79	55.11	51.94	63.81	68.25
Snowball_(bi-tri)	50.36	68.61	68.57	55.10	53.51	63.80	66.98
IL_unigram	58.62	65.69	75.32	73.86	61.63	76.61	79.30
IL_(uni-bi)	60.75	67.34	77.64	75.19	61.99	77.00	80.45
IL_(uni-bi-tri)	60.12	65.97	77.10	76.60	62.81	75.74	79.61
IL_bigram	61.55	67.19	74.72	75.60	60.57	75.89	80.68
IL_(bi-tri)	57.31	66.01	74.93	71.52	62.17	77.25	80.08
Lovins_unigram	62.60	64.84	78.13	69.76	65.78	72.02	79.88
Lovins_(uni-bi)	62.94	61.92	76.62	69.77	63.70	71.42	78.34
Lovins_(uni-bi-tri)	61.35	64.72	76.29	72.83	64.09	72.97	78.88
Lovins_bigram	61.74	61.31	77.26	73.52	64.15	73.94	79.13
Lovins_(bi-tri)	59.69	64.29	77.06	70.47	61.63	71.80	75.87
None_unigram	68.97	68.72	77.06	75.23	61.57	74.33	81.15
None _(uni-bi)	67.38	70.25	72.78	71.87	61.31	73.20	82.41
None _(uni-bi-tri)	68.00	71.41	76.08	74.78	61.22	75.04	84.53
None _bigram	65.25	72.41	79.40	74.27	58.62	73.66	82.83
None _(bi-tri)	63.78	69.41	72.36	75.50	56.49	75.52	83.62

Figure 3 depicts the overall average predictive performances of the seven classifiers based on four different stemming approaches in terms of accuracy and F-score. The classifiers using the None (raw data) stemming approach achieved the highest average accuracy and F-score values of 71.88 and 72.01 respectively, compared to other approaches. On the other hand, while LOVINS and IL stemmers performed similar results, the lowest average accuracy and F-score values of 61.79 and 61.86 respectively, have been achieved by the classifiers when the dataset stemmed by using the Snowball stemmer.

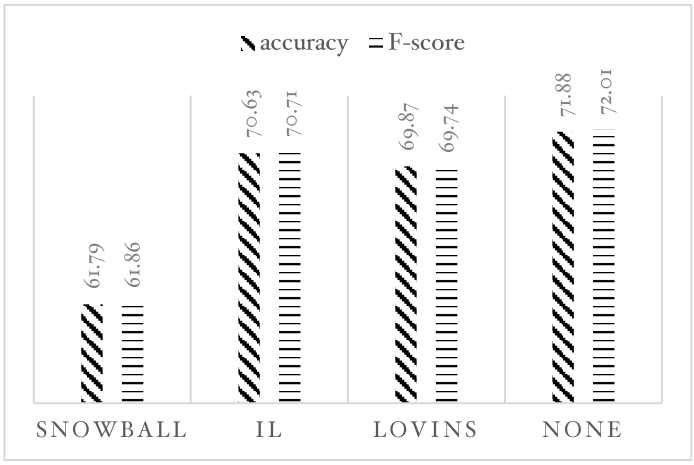

Figure 3. Average values of all cases based on four stemming approaches in terms of accuracy and F-score metrics.

Figure 4 displays the overall average accuracy and F-score values achieved by the classifiers when the three different ngram models (namely, unigram, bigram, trigram) and their combinations have been used to extract features. As it can be observed from the empirical results, the highest performances have been obtained by using all the three models together (uni-bi-tri) with the average accuracy and F-score values of 69.32 and 69.66, respectively. The second highest performances have been achieved by the combination of unigram and bigram models (uni-bi) with the results of 69 and 68.81. In contrast, the lowest accuracy and F-score values of 67.79 and 67.63 respectively, have been obtained by using the combination of bigram and trigram models (bi,tri).

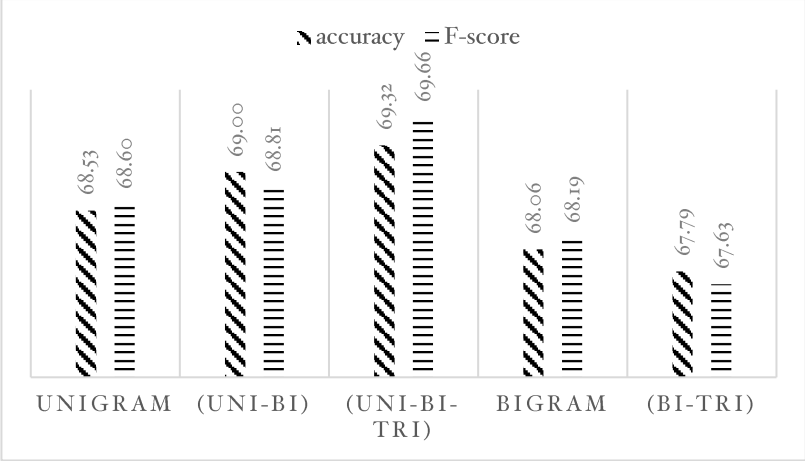

Figure 4. Average values of all cases based on five different ngram models in terms of accuracy and F-score metrics.

Figure 5 depicts the comparison of average predictive performances of the proposed classifier WMVE with six different machine learning algorithms in terms of accuracy and F-score values which are obtained from all cases within the empirical analysis. According to the results where accuracy and F-score values are slightly similar, the WMVE model outperformed other classifiers with an average accuracy value of 77.35. The second highest prediction performance has been achieved by REPTree with the accuracy value of 74.42. The third highest average accuracy value, which is 72.79, obtained by MV classifier. In contrast, KNN performed the lowest performance among all other classifiers.

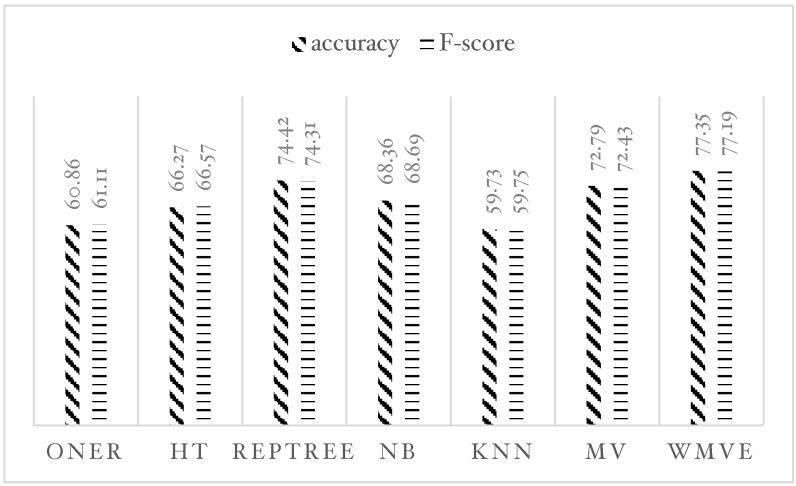

Figure 5. Average values of the classifiers in terms of accuracy and F-score metrics.

5. Discussion and Conclusion

In this paper, we presented a comprehensive sentiment analysis on Amazon customer reviews by proposing a novel weighted majority voting ensemble method utilizing five base classifiers which are Naïve Bayes, OneR, Hoefding Tree, REPTree, and KNN. We used a pre-collected Kaggle dataset which is composed of 9,641 reviews. In the empirical analysis phase, we compared the predictive performances in terms of accuracy, precision, recall and F-score metrics of WMVE and the six different machine learning classifiers (OneR, HT, REPTree, NB, KNN and MV) in conjunction with five ngram models (unigram, bigram, uni-bi, bi-tri, and uni-bi-tri). In addition, we evaluated the results of four various stemming approaches (namely, Snowball, IL, Lovins and None).

Regarding the overall predictive performances, the proposed model WMVE outperformed other classifiers by achieving the highest average accuracy and F-score values of 77.35 and 77.19 respectively. Besides, the empirical results show that the combination of three ngram models (uni-bi-tri) performed higher accurate performances compared to other ngram combinations. According to the results obtained by using different stemming approaches, we observed that using the raw dataset provided higher results compared the other stemming approaches.

It is clear that the proposed ensemble method with a special weighted majority voting mechanism performed higher scores compared to normal majority voting method and five different base classifiers. This paper's primary contributions can be succinctly outlined as follows: (i) it initiates with a concise overview of weighted majority voting methods, introduced to enhance the predictive performance of conventional ensemble learning approaches; (ii) it introduces a novel Weighted Majority Voting Ensemble (WMVE) that takes into account amplifying the impact of models correctly predicting outcomes based on the failure rate of other models; (iii) it showcases various experimental studies conducted on twenty-eight benchmark datasets, demonstrating that the proposed WMVE method generally yields superior classification outcomes compared to both the simple majority voting ensemble (MV) approach and

individual standard classification algorithms in terms of accuracy, precision, recall, and F-score. The classification algorithms employed in the experimental analysis include the REP decision tree, Hoefding Tree (HT), k-nearest neighbor (KNN), OneR, and naive Bayes (NB). These algorithms were chosen for their widespread popularity in this research.

For future work, we plan to use deep learning architectures (RNN, LSTM and CNN) as the base classifiers of the proposed ensemble model and compare the proposed model with transformer architectures such as Bidirectional Encoder Representations from Transformers (BERT) language model.

Ethics committee approval and conflict of interest statement

Ethics committee approval is not required for the prepared article.

There is no conflict of interest with any individual or institution in the prepared article

Author Contributions

Author 1 developed the classification method and conducted the experiments while author 2 completed necessary preprocessing operations on datasets and figured out an extensive literature review on text mining discipline. All of the authors contributed to the writing process of this manuscript.

References

[1] Basiri, E., Safarian, N., Farsani, E. 2019. A supervised framework for review spam detection in the Persian language, In: 2019 5th International Conference on Web Research (ICWR), 24-25 April, Tahran, Iran, 203-207.

[2] Juyal, P. 2022. Classification accuracy in sentiment analysis using hybrid and ensemble methods. 2022 IEEE World Conference on Applied Intelligence and Computing (AIC), 17-19 June, Sonbhadra, India, 583-587.

[3] Raza, N., Bharti, S., Ritika, M. 2023. Detecting the risk of Covid 19 Spread in near real-time using social media. International Journal of Emergency Management, vol. 18(2), 202-223. https://doi.org/: 10.1504/IJEM.2023.131940.

[4] Nona, N., Julien, K., Jenny, C., Patrick, R., Douglas, T. 2021. Ensemble of deep masked language models for effective named entity recognition in health and life science corpora, vol. 6, 689803. doi: 10.3389/frma.2021.689803

[5] McAuley, J., Leskovec, J. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews, https://doi.org/10.48550/arXiv.1303.4402

[6] Alharbi, N.M., Alghamdi, N.S., Alkhammash, E.H., Al Amri, J.F. 2021. Evaluation of sentiment analysis via word embedding and RNN variants for Amazon online reviews, Mathematical Problems in Engineering, vol. 2021, 1-10. https://doi.org/10.1155/2021/5536560

[7] Gondhi, N.K., Sharma, E., Alharbi, A.H., Verma, R., Shah. M.A. 2022, Efficient long short-term memory-based sentiment analysis of e-commerce reviews, Computational Intelligence and Neuroscience, vol. 2022,3464524. https://doi.org/10.1155/2022/3464524

[8] Dey, S., Wasif, S., Tonmoy, DS., Sultana, S., Sarkar, J., Dey, M. 2020. A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews, In: 2020 International Conference on Contemporary Computing and Applications, 05-07 February, Lucknow, India, 217-220.

[9] Khalid, M., Ashraf, I., Mehmood, A., Ullah, S., Ahmad M., Choi, GS. 2020, GBSVM: sentiment classification from unstructured reviews using ensemble classifier, Applied Sciences, vol. 10(8), 2788. https://doi.org/10.3390/app10082788

[10] Qorich, M., El Ouazzani, R. 2023, Text sentiment classification of Amazon reviews using word embeddings and convolutional neural networks, The Journal of Supercomputing, vol. 79, 11029–11054. https://doi.org/10.1007/s11227-023-05094-6

[11] Nandal, N., Tanwar, R., Pruthi, J. 2020, Machine learning based aspect level sentiment analysis for Amazon products, Spatial Information Research, vol. 28(5), 601-607. https://doi.org/10.1007/s41324-020-00320-2

[12] Alroobaea, R. 2022 Sentiment analysis on amazon product reviews using the recurrent neural network (rnn), International Journal of Advanced Computer Science and Applications, vol. 13(4), 5536560. https://doi.org/10.1155/2021/5536560

[13] Dogan, A, Birant, D. 2019. A weighted majority voting ensemble approach for classification. 2019 4th International Conference on Computer Science and Engineering (UBMK), 11-15 September, Samsun, Turkey, 1-6. doi: 10.1109/UBMK.2019.8907028

[14] Onan, A., Korukoglu, S., Bulut, H. 2016. Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications, vol. 57, 232-247. Doi: 10.1016/j.eswa.2016.03.045

[15] Bird, S., Loper, E. 2016. The natural language toolkit NLTK: The natural language toolkit, In: Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, March, 63–70.

[16] Frank, E., Hall, M.A., Witten, I.H. 2016. The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.