Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A High-Performance Model based on Ensembles for Twitter Sentiment Classification

R. Asgarnezhad[1], S.A. Monadjemi[2,*], M. Soltanaghaei[1]

[1]*Department of Computer Engineering, Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran.*

[2]Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran, and Senior Lecturer, School of continuing and lifelong education, National University of Singapore, Singapore, 119077.

Article Info

Article History:
Received 8 March 2019
Revised 05 June 2019
Accepted 28 November 2019

Keywords:
Text mining
Text classification
Machine learning
Ensemble method
Twitter

[*]Corresponding Author's Email Address:
monadjemi@eng.ui.ac.ir

Extended Abstract

Background and Objectives: Twitter Sentiment Classification is one of the most popular fields in information retrieval and text mining. Millions of people of the world intensity use social networks like Twitter. It supports users to publish tweets to tell what they are thinking about topics. There are numerous web sites built on the Internet presenting Twitter. The user can enter a sentiment target and seek for tweets containing positive, negative, or neutral opinions. This is remarkable for consumers to investigate the products before purchase automatically.

Methods: This paper suggests a model for sentiment classification. The goal of this model is to investigate what is the role of n-grams and sampling techniques in Sentiment Classification application using an ensemble method on Twitter datasets. Also, it examines both binary and multiple classifications, which are classified datasets into positive, negative, or neutral classes.

Results: Twitter Classification is an outstanding problem, which has very few free resources and not available due to modified authorization status. However, all Twitter datasets are not labeled and free, except for our applied dataset. We reveal that the combination of ensemble methods, sampling techniques, and n-grams can improve the accuracy of Twitter Sentiment Classification.

Conclusion: The results confirmed the superiority of the proposed model over state-of-the-art systems. The highest results obtained in terms of accuracy, precision, recall, and f-measure.

Introduction

With an increasing number of tweets over the Webs, tweets have interested more and more. There is a high interest in Sentiment Classification (SC) of tweets [1]-[9]. In many web sites, the user records an opinion, including positive, negative, or neutral sentiments [10]-[11]. Twitter classification is an outstanding problem, which has very few free resources and not available due to modified authorization status. However, all Twitter datasets are not labeled and free, except for our applied dataset.

Many kinds of research in Twitter Sentiment Classification (TSC) have converged on the usage of regular classifiers and machine learning-based classifiers [3], [7], [10]-[13]. The main problem in supervised techniques is the availability of labeled datasets [8]. We can only prepare a rare number of datasets for supervised models because manually collecting them is time-consuming. Also, a few studies [3] converged on the ensemble method. The

current authors in 2015 compared the validity of supervised and unsupervised approaches [14].
Yet, tweets are vaguer than other sentiment data like reviews [15].

Different challenges can be studied in TSC [3] concerning other datasets: classification accuracy, data sparsity problem, and neutral tweets. These cause to largest of tweets incorrectly classified. It revealed that Part of speech (POS) features were not helpful in the micro-blogging [1]. In the current study, we examine to define a way that increases classification performance.

In this article, we suggest a novel model, namely NEST, to improve TSC. Specifically, the boosting method, n-gram features, bootstrapping sampling, and Term Frequency–Inverse Document Frequency (TFIDF) weighting mechanism applied. The suggested model is novel, because it applies both binary and multiple datasets and combines n-gram, sampling techniques, and ensemble methods for TSC. We reveal that our model plays a vital role in the performance of the model. We produced multiple classifications containing positive, negative, or neutral tweets. We showed that the combination of TFIDF, sampling, and n-gram has a better result for both datasets. Also, we show that the usage of ensemble methods and combined with n-grams can increase the accuracy of TSC. Twitter-Sanders-Apple (TSA) datasets used in all experiments. The effectiveness of the suggested model compared with the methods in [4], [21], [26]-[27], [30], [32]. Our findings exposed that our features are well in two datasets. The obtained results presented in two experiments and validated that our model outperforms the existing methods on the datasets. The highest f-measure obtained 93.52% by our model on TSA2; whereas, Padmaja and Hegde [32] obtained 89.73%. Also, the best f-measure of the NSET achieved 89.64%; whereas, the highest f-measure in the literature obtained 81.25% by Pandey et al. [4] on the TSA3. It revealed that our model works better than the other methods based on genetic algorithm (GA) in [32] and cuckoo search in [4]. It also revealed that the NSET works better than the other methods based ensembles in [21], fuzzy rules in [26], [30], and supervised techniques in [27].

The innovations of this study indicated as follows:

- The model is of an ensemble nature
- Applying sampling technique and n-gram besides weighting mechanism for improving classification efficiency
- Providing the boosting selector in conjunction with the popular classifier
- Choosing the best features based on the feature selection stage and two error indices
- Employing both two and three classes datasets on Twitter

The rest of this article organized as follows: Sec. 2 and Sec. 3 presents available techniques and related works, respectively. The proposed model is shown in Sec 4 and evaluated in Sec. 5. Finally, the conclusion and future works display in Sec. 6.

Available Machine Learning Techniques

The machine learning (ML) techniques for text classification algorithms, like maximum entropy (ME), naive Bayes (NB), and support vector machines (SVM), have achieved peerless success in SC text categorization, and these classifiers provided feasibility in their tasks. On reviewing the experiment dataset, the results of SVM was virtually better than other ML techniques. Hence we in our model used it to improve classification performance. Here, the summarization of some of the popular ML techniques in this context is of concern.

The ML approaches applied supervised, unsupervised, and semi-supervised methods and employed linguistic features. The lexicon-based approaches divided into corpus-based and dictionary-based approaches. The main advantage of them is to support in determining domain and context-specific opinion words using a domain corpus. In lexicon-based approaches, a document divides by aggregating the sentiment orientation of all available words. A document with more positive words classified as positive; whereas, the document with more negative words categorized as negative. Hybrid approaches combined the advantages of both approaches to improve the performance of SC.

Explanations of some classifiers are of concern herein.

A. Naïve Bayes

NB obtained reasonable accuracy. It is simple and assumed independent features.

Also, it mainly used when the size of the training set is not vast. Here, (1) is applied [16] to calculate the probability of event A in column A, provided that class C holds.

$$P(K = A \mid C) = \frac{1}{\sqrt{2\pi\sigma^2_{K=C}}} e^{-\frac{A-\mu_{K=C}}{2\sigma^2_{K=C}}} \quad (1)$$

where, $\mu k = c$ is the column K mean, while the row belongs to the class C and $\sigma^2_{k=c}$ is the variance of the kth therein, and no input classification is required. An example presented to explain the Bayes Continuous Decider, where, there exist four features with positive or negative classes.

B. Maximum Entropy

Unlike NB, ME is assumed dependent features [17]. This technique estimates $P(c \mid d)$ in

$$P_{ME}(c \mid d) = \frac{1}{Z(d)} e^{\sum_i \lambda_{i,c} F_{i,c}(d,c)} \quad (2)$$

where, Z(d) is a normalization function and $F_{i,c}$ is a function for feature F_i and class *c*, as in

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \quad and \quad c' = c \\ 0 & otherwise \end{cases} \tag{3}$$

C. Neural Network

Here, a description of the perceptron classifier is of concern: If *m* is the count of the selected features and the dataset is named P, each user named P_i would have been assigned the *m* features, and if any connection attribute *x* is considered, there are variables x_1 to x_m for each connection. These inputs are samples of the training network. It is the training method with a supervisor because the network is trained through samples with the correct output (Fig. 1).

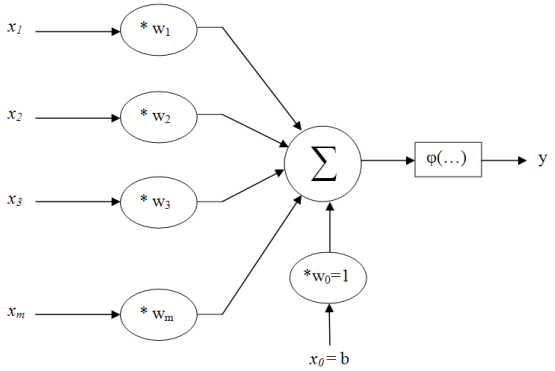

Fig. 1: Single-layer perceptron [18].

D. Support Vector Machine

In this structure, first, the attribute table converts into a set of data points $\{(x_1,c_1),(x_2,c_2),...,(x_n,c_n)\}$, and then, these divide into two classes $c_i = \{-1,1\}$. Each x_i is a p-dimensional vector of real numbers, which are the same properties extracted from the previous step.

Linear classification methods try to separate data by constructing a hyperplane, which is a linear equation). The SVM classification method, which is one of the linear classification methods, finds the best hyperplane that separates data from two classes with maximum margin. A picture of a data set belonging to two classes, which selects the best hyperplane for separating them exposed in Fig. 2. In this form, the data is two-dimensional, that is, each data consists of only two variables [19].

Here explains in detail how to produce a separator hyperplane. An accurate picture of how the separator hyperplane produced through the SVM exposed in Fig. 3.

First, consider a convex hull around the points of each class. In Fig. 3, the convex hull drawn around the points related to class -1 and class +1. Line P is the line that shows the closest distance between two convex hulls. *h*, which is the separating hyperplane, is a line that splits P and is vertical to it. The *b* is the width of the source for the hyperplane with the maximum separation limit. If *b* ignored, the solutions are the only hyperplane that goes beyond the source. The vertical distance of the hyperplane to the source achieved through dividing the absolute value of the parameter *b* by the length *w*.

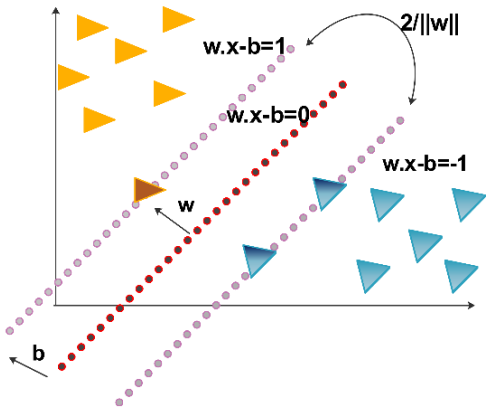

Fig. 2: Hyperplane with maximum separator boundary with separating boundaries for classification.

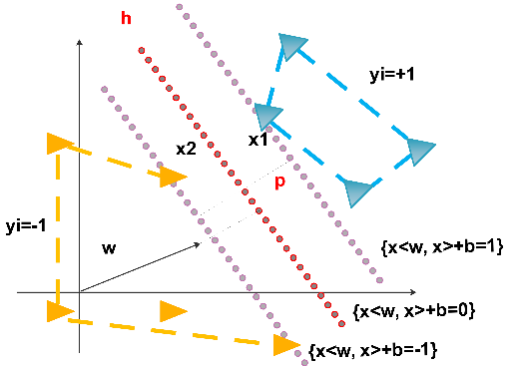

Fig. 3: How to build a separating hyperplane between two data classes in two-dimensional space.

The basic idea is to choose the proper separator. It refers to the separator that is farthest from the neighboring points on both floors. This answer has the highest boundary with points on two different floors and can be bounded by two parallel hyperplanes that pass through at least one of the floor points. These vectors are named support vectors. The mathematical equations for these two parallel hyperplanes are of concern (5) and (6).

$$w.x - b = 1 \tag{4}$$

$$w.x - b = -1 \tag{5}$$

It is remarkable to remark that if the training data are linearly separable, the two boundary hyperplanes can be chosen in such a way that there is no data between them, and then, the distance between the two parallel hyperplanes can be maximized. Applying geometric theorems, the distance between the two hyperplanes is $|\frac{w}{2}|$, So you have to $|w|$ Minimized. It is also necessary to

prevent data points from being placed within the boundary, for which a mathematical constraint added to the formal definition. For each *i*, it ensured through employing the following constraints that no point placed on the boundary. For data related to the first and second floors, (7) and (8) are of concern, respectively.

$$w.x_i - b \geq 1 \quad (7)$$

$$w.x_i - b \leq -1 \quad (8)$$

The following constraint can be shown as follows.

$$c_i(w.x_i - b) \leq 1 \qquad 1 \leq i \leq n \quad (9)$$

Related Work

SC has attracted a great deal of attention in recent years. A large number of methods proposed for improving classification performance. These methods differ from each other in the way the architecture of the classifier, algorithm parameters, or preprocessing methods. Here, the summarization of some of the existing articles on the TSA datasets are of concern:

In 2009, Liu et al. suggested an ESLAM model. They try to train a language model based on manually labeled data [20].

Expressed ensemble techniques are effective for SC of feature sets and classification algorithms. For example, in 2013, Hassan et al. proposed an ensemble framework in [21] and used a combination of unigrams and bigrams, POS, and semantic features derived from WordNet (WN) and SentiWordNet (SWN). Authors applied several base learners like NN, Random Tree (RT), NB, Bayes Net, Logistic Regression (LR), and SVM. Despite using the bootstrap model and several classifiers, their framework was not more effective than our approach. They obtained 76.30% of accuracy. Unfortunately, their different datasets are not available. In 2015, Lima et al. [22] introduced a polarity analysis framework for Twitter, which follows ML approaches. They utilized four datasets to estimate the performance of their framework. Additionally, they employed five kinds of classifiers like NB, SVM, Decision Trees (DT), and Nearest Neighbors (KNN).

In 2017, Keshavarz and Abadeh [23] combined both corpora and lexicon approaches. For this goal, they produced lexicons from the text. Also, they applied a novel GA to solve the SC problem. Adaptive sentiment lexicons generated by the algorithm to choose the best features. Their experiments conducted on six datasets. Also, Bala [24] used supervised and unsupervised techniques. The results obtained on three labeled datasets. Also, the author conducted a feature selection using a GA to verify results. The experiments reveal that the obtained results via supervised techniques are different on datasets. After the preprocessing stage, the document term matrix produced using unigram and bigram. Next, features extracted and supervised learning algorithms like NB, SVM, and DT applied to the datasets. Also, Pandey et al. [4] proposed a novel clustering method using k-means and cuckoo search in 2017. They achieved an accuracy of 81.4 and 82.20% for TSA2 and TSA3, respectively.

In 2018, Haider et al. [25] investigated the impact of adverbs for SC. Also, Trupthi et al. [26] investigated the effective topic modeling methodology Latent Dirichlet Allocation (LDA) to extract the keywords in a clustering manner. Next, they applied the keywords using the Possibilistic fuzzy c-means approach for twitter sentiment analysis.

The present researchers in [27] proposed a model named SFT for TSC in 2018. The goal of our model was to investigate the role of weighting feature techniques in SC using supervised methods on the Twitter data set. The applied classifier in the current article is based on the SFT model in our previous article.

Abdolahi and Zahedi in [28] introduced a method to consolidate the external word correlation knowledge into short and long stories in both local and global coherence in 2018. Using the effect of combined word2vec vectors, they confirm that their proposed method is free of the language and its semantic concepts. They received 87.03% of accuracy. Behravan et al. [29] in 2018 suggested a new clustering method for big datasets using Particle Swarm Optimization (PSO) algorithm. Their proposed method was a two-stage algorithm: (1) search the solution space for a proper number of clusters and (2) search to find the position of the centroids.

In 2018, Vashishtha and Susan in [30] estimated the sentiment of social media posts using a novel set of fuzzy rules. Their system combines Natural Language Processing techniques and Word Sense Disambiguation using nine fuzzy rule-based systems. They reached 59.7, 58.9, and 68.6% of precision, recall, and f1-score on the TSA3 datasets, respectively.

In 2019, Tripathi et al. [31] suggested a novel Map-Reduce based K-means to cluster the large scale data. Also, Padmaja and Hegde [32] proposed a system consists of three phases; data collection, preprocessing, and classification of sentiments. In the third phase, a hybrid classifier applied to classify the twitter sentiment classes.

In 2020, Abbas et al. [33] offered a classification model with four classifiers, and varying techniques consist of NB, DT, multilayer perceptron, and LR to form a single ensemble classifier. They gained an accuracy of 82.2% on Twitter. Also, Jiang et al. [34] develop a novel NN-based model, namely MAN, to conduct the aspect-level SC tasks.

In 2020, Naseem et al. [35] shown a transformer-based method for SA and applied deep intelligent contextual embedding to heighten the quality of tweets by removing noise. They also employed the bidirectional Long Short Term Memory (LSTM) network to define the sentiment of a tweet. They reached an accuracy of 96.2% on airline datasets.

In 2020, Samad et al. [36] studied the effect of seven text processing scenarios on Twitter. Their experiments revealed negative effects on SC of two common text processing steps: 1) stop word removal; 2) averaging of word vectors to represent individual tweets. Word selection from context-driven word embedding showed that only the ten most important words in Tweets cumulatively produce over 98% of the maximum accuracy. Also, Nemattolahzadeh et al. [37] suggested a method to utilize experimental data for identifying the influence network between individuals in social networks. Their method was based on convex optimization and could identify interaction patterns accurately. The three models were the most comprehensive and vastly models in the literature considered.

In 2020, Sharma and Jain in [38] presented a study on Twitter sentiment analysis where tweets collected and sentiments behind the tweet assessed using various ML techniques. They extracted data from twitter, and text preprocessing and feature extraction employed to the textual data. Correlation-based attribute selection methods applied and ML classifiers consist of SVM, NB, Random Forest, Meta classifier, and LR analyzed to confirm which classifier gives better results. They obtained an accuracy of 88.2% on the Cambridge Analytica dataset.

The NSET Model

The NSET model proposed in five stages: (1) preprocessing; (2) sampling; (3) weighting mechanism; (4) feature selection; (5) classification; and (6) performance evaluation. Fig. 4 tabulated the NSET model in detail.

Three main contributions of the NSET decorated in orange color.

In preprocessing, the n-grams applied to handle important relations. Next, a bootstrapping sampling performed to boost accuracy. After performing the TFIDF, classification methods run on test datasets. The AdaBoost method through the 10-fold cross-validation scheme on the dataset adjusted. In our model, shuffled and stratified samplings applied. In all experiments, results in terms of 10-fold cross-validation obtained.

The goal of this article is to study the role of n-grams and sampling using an ensemble method. Here, the stages of the model described.

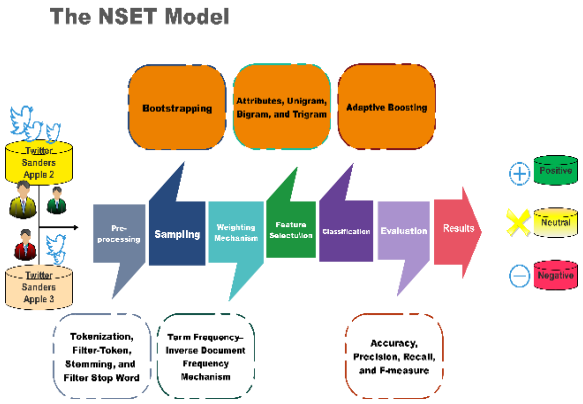

Fig. 4: The steps of the NSET for Twitter Sentiment Classification.

A. Preprocessing

Preprocessing stages including Tokenization, Filtering Token, Stemming, Filtering Stop Word, and N-grams. First, useless characters and words tokenized. Each tweet convert in a sequence of tokens and filters based on their length. The root of each word found through stemming. According to the stop word list, the stop words in each tweet eliminated and filtered.

B. Sampling

Here, attribute subset selection techniques used to improve the classification performance. Bootstrapping technique applied to obtain higher accuracy in our NSET model. This type of sampling applied sampling with replacement. So, the sample may not have all unique examples. Once an example selects, it remained a candidate for selection and can choose again. Additionally, it may generate a sample that is greater in size than the original dataset [39]. When a tuple selected, it has equal probability to select and add to the training set again.

C. Weighting Mechanism

TFIDF weighing mechanism used to produce word vector. It consists of two ratings, regularity and inverse regularity of phrase. Inverse document frequency investigated by splitting the number of records. TFIDF mechanism defined as in

$$TFIDF = TF \,.\, Log\,(N / F_t) \tag{10}$$

TF is the frequency of word t in document d, N is the number of documents, and Ft is the number of documents, including word t. It did not assign high scores to frequent words [40].

D. Feature Selection Methods

Feature selection methods extract a subset of features from all possible lists of features in the dataset to present the prediction results. These methods work at two levels: first, the selection of the subset of attributes through an attribute evaluator algorithm, and second is

the evaluation of the search heuristics via a search algorithm. Feature selection can be done in three ways:

• Evaluating the performance of the set of attributes by using a specific classifier, namely wrapper method.

• Selecting attributes as a filter in the preprocessing phase of data analytics.

• Selecting a set of attributes as a unigram, bigram, and trigram in the preprocessing phase of data analytics.

E. Classification

Ensemble methods apply multiple models to obtain better predictive performance than could be obtained from any of the constituent models. Ensembles may become more flexible in the functions. However, some ensemble techniques, especially bagging, tend to reduce problems related to over-fitting of the training data. Empirically, ensembles tend to yield better results when there is significant diversity among the models. Boosting is an ensemble method that can be used in conjunction with many other learning algorithms to improve their performance.

AdaBoost is a nested method and tries to build a better model using the learner provided in its subprocess. AdaBoost is a meta-algorithm and can be used in conjunction with many other learning algorithms to improve their performance. AdaBoost is sensitive to noisy data and outliers. However, it can be less susceptible to the overfitting problem than most learning algorithms.

The classifiers it uses can be weak, but as long as their performance is not random, they will improve the final model. To sum up, we used AdaBoost in conjunction with the SVM classifier [41].

Here, the pseudo-code of our model expressed as:

Pseudo-code for the NSET model

```
while (Website is Online) do
TW=Get-T (Tweets)
for each tweet in TW do
    Tokenizing, Splitting, Filtering, Stemming,
   Omitting Stop words, and Generating N-gram
 end for
W=Pre-process (a set of words)
W-s=Sample (W)
for each set of words in W-s do
    Constructing a word vector via TFIDF schema
 end for
WV=Get-TFIDF (a set of weights)
F=Feature-Selection (WV)
for each word vector in F do
    Applying the bootstrapping, storing the results
    Applying the SVM classifier
 end for
Best-F=SVM-Classifier (F)
Model=Ensemble (Best-F)
Per=Evaluation (Model)
end while
```

Here, the pseudo-code of our ensemble expressed as:

Pseudo-code for the Ensemble

```
for each classifier do
    W=Weigh-Vote (F)
    C=Predict (W)
  Add W to weight for the class
 end for
return the class with the biggest weight
```

Results and Discussions

Here, measures for evaluating SC introduced. P and N are the numbers of positive and negative tuples. TP refers to the positive tuples that correctly labeled by the classifier. TN refers to the number of true negatives. FP is the negative tuples that incorrectly labeled as positive. FN is the positive tuples that mislabeled as negative. Accuracy is the sum of actual tuples that classified TP and the number of TN relative to the total number of classified instances. Precision stated as the percentage of tuples that labeled as positive and actual. Recall refers to the percentage of tuples that labeled positive. F-measure combines precision and recall into a single measure [39]. F-measure comes from a weighted harmonic mean of precision and recall. Also, mean absolute error (MAE) and root absolute error (RAE) for error evaluation employed. These measures computed in Eq. (11) to (16) [42]-[43].

$$Accuracy = \frac{TP + TN}{P + N} \tag{11}$$

$$Precision\,(P) = \frac{TP}{TP + FP} \tag{12}$$

$$Recall\,(R) = \frac{TP}{P} \tag{13}$$

$$F - measure = \frac{2PR}{P + R} \tag{14}$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{15}$$

$$RAE = \frac{\sum_{i=1}^{n} |x_i - \overline{x_i}|}{\sum_{i=1}^{n} |y_i - \overline{y_i}|} \tag{16}$$

F. Dataset

Two datasets prepared by Niek Sanders applied. However, the datasets used in other studies are not

openly available, except for Sanders. You can found it in this address, http://www.sananalytics.com. The detailed information of two corpora shown in Table 1.

Table 1: A description of the used datasets

Datasets	Number of instances	Number of instances in classes		
		Positive	Negative	Neutral
TSA2	479	163	316	-
TSA3	988	163	316	509

Dataset 1: This dataset is a subset of TSA and consists of 479 tweets. There are 163 positive and 316 negative tweets in the given dataset.

Dataset 2: This dataset is also a subset of TSA and contains 988 tweets. It has three classes having 163 positive, 316 negative, and 509 neutral tweets.

G. Experiments

To achieve state-of-the-art results R implementation applied to conduct experiments on the sanders dataset. The accuracy and efficiency of the NSET model examined. The experiments try to evaluate the effectiveness of n-grams and sampling. Moreover, the impact of the ensemble method and the combination of preprocessing techniques estimated.

Assumptions**:** The default setting chosen in all experiments. For classification, a supervised method of SVM as a base learner used to receive the highest performance and combine supervised and ensemble methods. In our previous work, experiments showed that SVM is one of the best classifiers on the TSA dataset. The linear kernel and the value for the parameter epsilon=1.0 for LibSVM (C-SVC) determined by cross-validation. All related parameters set optimal. For cross-validation, 10-fold using shuffled used in all experiments.

Experiment I. The first experiment investigated the effect of the TFIDF mechanism and n-grams on the evaluation metrics using AdaBoost and sampling on TSA2. We used Term Frequency and TFIDF mechanisms in the primary preprocessing stage. However, it reveals that the Term Frequency mechanism cannot improve performance. The TFIDF weighting mechanism was useful. Hence, the TFIDF mechanism considered both experiments. We showed that 10-fold cross-validation with shuffled often is better than a stratified one. Moreover, the data using bootstrapping sampling reduced. The obtained results obtained shown in Table 2.

The highest results in each column of the table marked as bolded text. The highest accuracy highlighted at 90.61%. Also, this result achieved when used bigrams. Bigrams provide a good balance among unigrams and an ability to obtain the sentiment expression patterns. However, SVM confuses when trigrams used. It found that the highest f-measure is 93.52%, which belongs to bigrams.

Experiment II. The second experiment investigated the effect of the TFIDF mechanism and n-grams on the evaluation metrics using AdaBoost and sampling on the TSA3.

We showed that 40-fold cross-validation with stratified is better than shuffled one for this dataset. Therefore, 10-fold cross-validation using stratified used in this experiment. Also, the data using bootstrapping sampling reduced. The obtained results showed in Table 3.

The highest accuracy highlighted 87.65%, which belongs to the unigram feature. However, SVM confuses when n-gram with higher levels applied. The highest f-measure obtained at 89.64%.

Table 2: The performance of the NSET on the TSA2 (%)

N	ACT / PRE	Confusion matrix		Results			
		POS	NEG	P	R	F	A
1	POS	113	11	91.13	75.84	82.78	90.19
	NEG	36	319	89.86	96.67	93.14	
2	POS	109	5	95.61	73.15	82.89	90.61
	NEG	40	325	89.04	98.48	93.52	
3	POS	105	5	95.45	70.47	81.08	89.77
	NEG	44	325	88.08	98.48	92.99	

Note: ACT= Actual, POS=Positive, NEG=Negative, PRE=Prediction, P=Precision, R=Recall, F=F-measure, A=Accuracy

As shown in Fig. 5, the highest accuracy obtained when bigrams used. But this matter was not verified for TSA3.

Fig. 6 compares the evaluation metrics for the NSET on the datasets. Also, the error indices of two experiments show that the obtained results can be good enough (Table 4).

For TSA2, all results are higher than those of TSA3. Besides, the values of precision are above 95% for both datasets.

We showed that the NSET make a batter for binary classification according to these datasets. Nonetheless, this model outperforms the existing methods for multiple classification tasks.

Table 3: The performance of the NSET on the TSA3 (%)

N	ACT PRE	Confusion matrix POS	NEG	NEU	Results P	R	F	A
1	POS	121	2	4	95.28	70.35	80.94	
	NEG	4	243	19	91.35	83.51	87.25	87.65
	NEU	47	46	502	84.37	95.62	89.64	
2	POS	119	3	8	91.54	69.19	78.81	
	NEG	6	239	18	90.87	82.13	86.28	86.74
	NEU	47	49	499	83.87	95.05	89.11	
3	POS	118	3	5	93.65	68.60	79.19	
	NEG	4	231	15	92.40	79.38	85.40	86.43
	NEU	50	57	505	82.52	96.19	88.83	

Note: ACT= Actual, POS=Positive, NEG=Negative, NEU= Neutral, PRE=Prediction, P=Precision, R=Recall, F=F-measure, A=Accuracy

Table 4: Error comparison

Experiment N0.	MAE	RAE
I	0.0876	0.0978
II	0.0977	0.0998

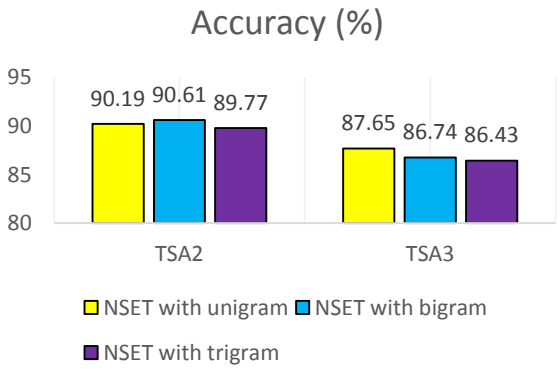

Fig. 5: The highest accuracy for the NSET based on n-grams.

H. Discussion

Here, the effects of n-grams and ensemble investigated on the TSA. The effectiveness of the suggested model examined the two datasets and compared them with the methods in Trupthi et al. [26], our previous model [27], Padmaja and Hegde [32], Vashishtha and Susan in [30], Hassan et al. [21], and Pandey et al. [4].

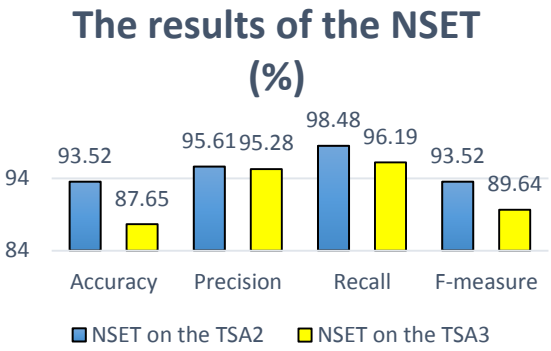

Fig. 6: The highest results for the NSET on two datasets.

The obtained results presented in two experiments and validated that the NSET outperforms the existing methods on the datasets, except for the achieved accuracy on the TSA2. Fig. 7 illustrates the performance of the NSET model and others in terms of accuracy. The highest accuracy gained 92.78% by Padmaja and Hegde [32] on the TSA2 and 82.2% by Pandey et al. [4]. In the TSA3, whereas the NSET received 90.61 and 87.65% for TSA2 and TSA3, respectively. However, we examine more alternatives to improve accuracy.

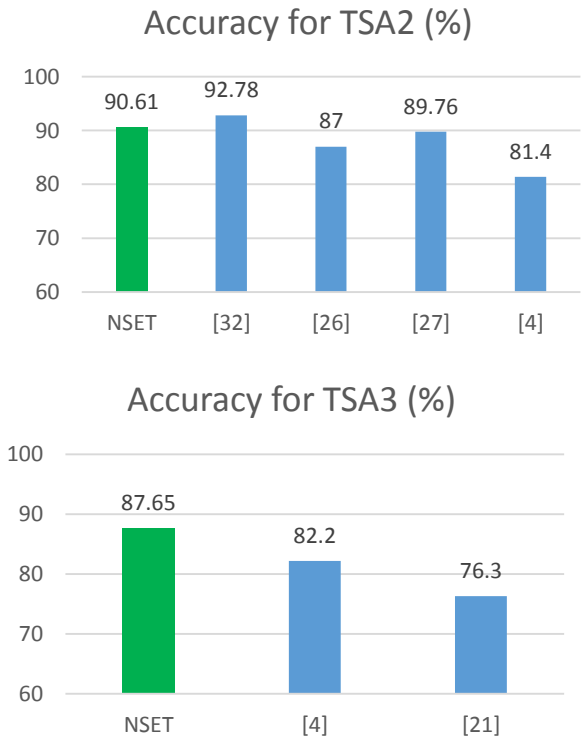

Fig. 7: The comparison of accuracy among the NSET and others.

Fig. 8 illustrates the performance of the NSET model and others in terms of precision. The best precision in

the literature achieved 90.4% by our previous model [27] on the TSA2 and 82.34% by Pandey et al. [4] on the TSA3. The highest precision gained 95.25% by the NSET, whereas Padmaja and Hegde [32] obtained 85.2%. It was approximately 10% higher than that one. We show that bigram features can work well using our model on the TSA2, but with the increasing to trigram, the embedded SVM in our mode confuse and it cannot improve more. Besides, the increase of n in the model causes few improvements in some cases of the TSA3. It also revealed that our model is better than the proposed model by Pandey et al. [4] through a cuckoo search on the TSA3. It finds that an uncomplicated and accurate model can be good enough in this context.

Fig. 9 illustrates the performance of the NSET model and others in terms of recall. On the other hand, the best recall of the NSET gained 98.48% on the TSA2, but Padmaja and Hegde [32] obtained 90.05%. This value for Trupthi et al.

[26] and our previous model [27] achieved 89.06 and 85.4%, respectively. It is also shown that the recall of our model received 96.19% on the TSA3; whereas, Pandey et al. [4] obtained 80.16%, an increase of approximately 16%. These differences revealed that the best criteria can be the f-measure.

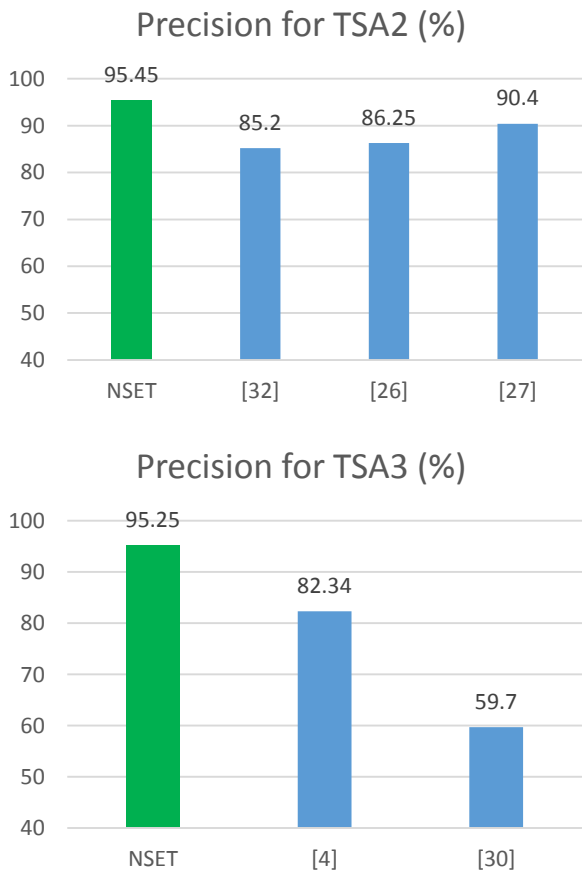

Fig. 8: The comparison of precision among the NSET and others.

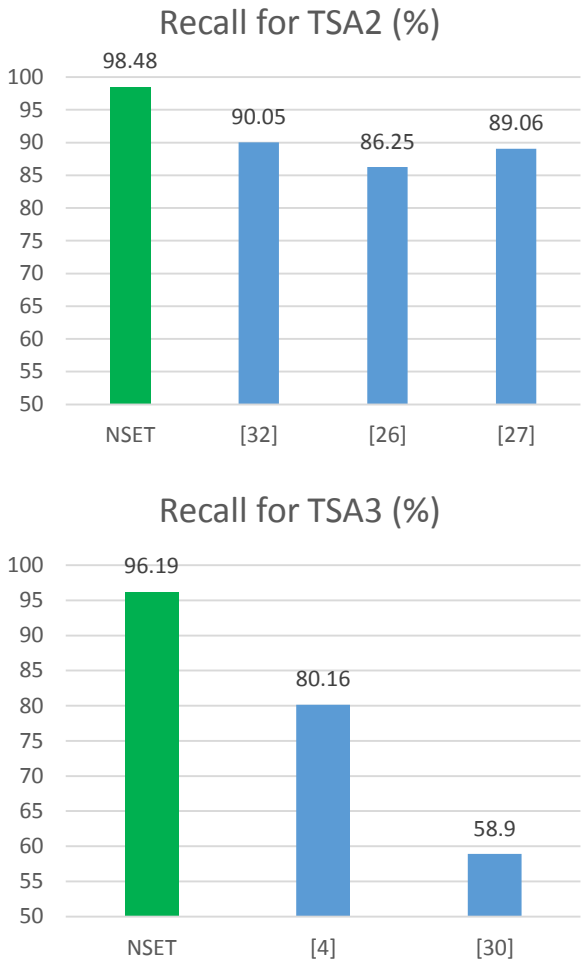

Fig. 9: The comparison of recall among the NSET and others.

It reflects the highest f-measure gains in both datasets. The f-measure for the comparison works is of concern in Fig. 10.

It is clear from the comparison that the NSET shows better accuracy for classification. It is a notable difference between the models. The highest f-measure reached 93.52% by our model on TSA2; whereas, Padmaja and Hegde [32] obtained 89.73%, approximately 4% more. The best f-measure of the NSET gained 89.64%; whereas, the highest f-measure in the literature received 81.25% by Pandey et al. [4] on the TSA3.

The results showed that the NSET is more accurate than its predecessors. It showed that the combination of TFIDF, sampling, and n-gram is a good alternative for both used datasets.As a limitation in the used dataset, the tweets are very short in the TSA3. Therefore, the obtained results have not a significant improvement, and these are the same approximately in this dataset. For this reason, we want to work more in the preprocessing stage to choose the best feature in this dataset.

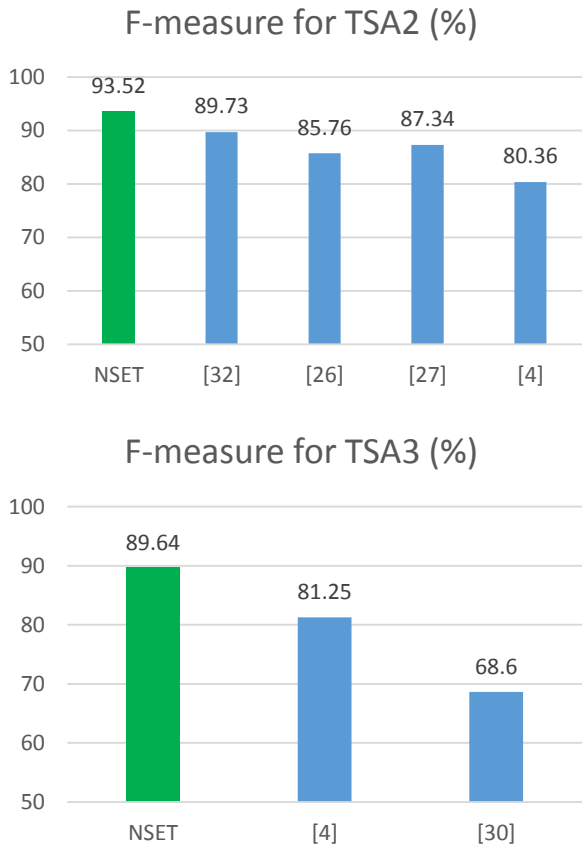

Fig. 10: The comparison of f-measure among the NSET and others.

Conclusion

The NSET suggests a model for both binary and multiple classifications, which classify the dataset into positive, negative, or neutral classes. Compared to the existing studies on Twitter SC which, either depends on sophisticated features or complicated learning procedure, the NSET is more simple and straightforward. The effect of n-grams using the ensemble method and sampling technique on the Twitter datasets is investigated. The highest f-measure achieved 93.52%, which belongs to bigrams. Bigram features construct the relationship of words to improved results. Experimental results demonstrated that the present model outperforms the existing methods based on two experiments on the datasets. Maximum precision obtained 95.45%, an increase of 10%. The NSET is very redeeming in comparison to others since it applied more related words using n-grams and sampling techniques.

The results revealed that the NSET is more accurate than its predecessors. It is also shown that the combination of TFIDF, sampling, and n-gram is a good alternative for both datasets. Our findings exposed that bigram features are well only in the TSA2; whereas, unigrams achieved the best performance for the TSA3. It revealed that our model works better than the other methods based GA in [32], cuckoo search in[4], ensembles in [21], fuzzy rules in [26], [30], and supervised techniques in [27].

We believe that performance can still be improved. As future work, we aim to study the use of heuristic algorithms as a way to improve feature selection and reduce the feature.

Authors contributions

R. Asgarnezhad designed the experiments, carried out the data analysis, interpreted the results and wrote the manuscript. S. A. Monadjemi corrected the proofing the article. Soltanaghaei supported the article.

Acknowledgment

We thank the editor and all anonymous reviewers.

Conflict of Interest

The authors declare that they have no conflict of interest.

Abbreviations

TFIDF	Term Frequency-Inverse Document Frequency
$\mu k = c$	The column K mean and the row belongs to the class C
$\sigma^2_{k=c}$	The variance of the kth
Z(d)	A normalization function
$F_{i,c}$	A function for feature F_i and class *c*
m	The count of the selected features
P	The dataset
$\{(x_1,c_1),(x_2,c_2),...,(x_n,c_n)\}$	A set of data points
x_i	A p-dimensional vector of real numbers
TF	The frequency of word *t* in document *d*
N	The number of documents
Ft	The number of documents including word *t*
MAE	Mean Absolute Error
RAE	Absolute Error

References

[1] E. Kouloumpis, T. Wilson, J. D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in Proc. Fifth International AAAI conf. on weblogs and social media, 2011: 538-541, 2011.

[2] F. H. Khan, S. Bashir, U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, 57: 245-257, 2014.

[3] N. F. Da Silva, E. R. Hruschka, E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, 66: 170-179, 2014.

[4] A. C. Pandey, D. S. Rajpoot, M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," Information Processing & Management, 53: 764-779, 2017.

[5] H. Saif, M. Fernández, Y. He, H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," in Proc. Ninth International Conf. on Language Resources and Evaluation, 2014: 810–817, 2014.

[6] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, B. Qin, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," in Proc. The 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 1555-1565, 2014.

[7] B. Besbinar, D. Sarigiannis, P. Smeros, "Tweet Sentiment Classification," Lausanne, 2014.

[8] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, "Ranked wordnet graph for sentiment polarity classification in twitter," Computer Speech & Language, 28: 93-107, 2014.

[9] D.-T. Vo, Y. Zhang, "Target-Dependent Twitter Sentiment Classification with Rich Automatic Features," in Proc. IJCAI,: 1347-1353, 2015.

[10] A. Go, R. Bhayani, L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford, 1, :1-6, 2009.

[11] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, "Target-dependent twitter sentiment classification," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- 1: 151-160, 2011.

[12] H. Saif, Y. He, H. Alani, "Alleviating data sparsity for twitter sentiment analysis," in Proc. the 21st International Conference on theWorld Wide Web,: 2–9, 2012.

[13] L. Chen, W. Wang, M. Nagarajan, S. Wang, A. P. Sheth, "Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter," ICWSM, 2: 50-57, 2012.

[14] R. Asgarnezhad, K. Mohebbi, "A Comparative Classification of Approaches and Applications in Opinion Mining," International Academic Journal of Science and Engineering, 2(1): 68-80, 2015.

[15] R. Asgarnezhad, S. A. Monadjemi, M. Soltanaghaei," FAHPBEP: A fuzzy Analytic Hierarchy Process framework in text classification," accepted in Majlesi Journal of Electrical Engineering, 14(3), 2020.

[16] J. Han, "MichelineKamber. Data mining: concepts and techniques," Morgan Kaufmann Publishers–An Imprint of Elsevier, 500: 105-150, 2006.

[17] S. R. Ahmad, M. Z. M. Rodzi, N. S. S. Nurhafizeh, M. M. Yusop, S. Ismail, "A Review of Feature Selection and Sentiment Analysis Technique in Issues of Propaganda," International Journal of Advanced Computer Science and Applications, 10(11): 240-245, 2019.

[18] J. J. Shynk, "Performance surfaces of a single-layer perceptron," IEEE Transactions on Neural Networks, 1: 268-274, 1990.

[19] D. Michie, D. J. Spiegelhalter, and C. Taylor, "Machine learning," Neural and Statistical Classification, 13: 1-298, 1994.

[20] K.-L. Liu, W.-J. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in Proc. The Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012: 1678–1684, 2012.

[21] A. Hassan, A. Abbasi, and D. Zeng, "Twitter sentiment analysis: A bootstrap ensemble framework," in Proc. International Conference on Social Computing,: 357-364, 2013.

[22] A. C. E. Lima, L. N. de Castro, and J. M. Corchado, "A polarity analysis framework for Twitter messages," Applied Mathematics and Computation, 270: 756-767, 2015.

[23] H. Keshavarz and M. S. Abadeh, "ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs," Knowledge-Based Systems, 122: 1-16, 2017.

[24] M. Bala, "Sentiment Classification Using Supervised and Unsupervised Approach," International Journal on Future Revolution in Computer Science & Communication Engineering, 3(11): 573-577, 2017.

[25] S. Haider, M. Tanvir Afzal, M. Asif, H. Maurer, A. Ahmad, A. Abuarqoub, "Impact analysis of adverbs for sentiment classification on Twitter product reviews," Concurrency and Computation: Practice and Experience: 1-15, 2018.

[26] M. Trupthi, S. Pabboju, G. Narsimha, "Possibilistic fuzzy C-means topic modelling for twitter sentiment analysis," International Journal of Intelligent Engineering and Systems, 11: 100-108, 2018.

[27] R. Asgarnezhad, S. A. Monadjemi, M. Soltanaghaei, A. Bagheri, "SFT: A model for sentiment classification using supervised methods in Twitter," Journal of Theoretical & Applied Information Technology, 96(8): 2242-2251, 2018.

[28] M. Abdolahi, M. Zahedi, "A new model for text coherence evaluation using statistical characteristics," Journal of Electrical and Computer Engineering Innovations (JECEI), 6: 15-24, 2018.

[29] I. Behravan, S. H. Zahiri, S. M. Razavi, R. Trasarti, "Clustering a Big Mobility Dataset Using an Automatic Swarm Intelligence-Based Clustering Method," Journal of Electrical and Computer Engineering Innovations (JECEI), 6: 243-262, 2018.

[30] S. Vashishtha, S. Susan, "Fuzzy rule based unsupervised sentiment analysis from social media posts," Expert Systems with Applications, 138: 1-15, 2019.

[31] A. K. Tripathi, K. Sharma, M. Bala, "Parallel hybrid bbo search method for twitter sentiment analysis of large scale datasets using mapreduce," International Journal of Information Security and Privacy (IJISP), 13: 106-122, 2019.

[32] K. Padmaja, N. P. Hegde, "Twitter sentiment analysis using adaptive neuro-fuzzy inference system with genetic algorithm," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019: 498-503, 2019.

[33] A. K. Abbas, A. K. Salih, H. A. Hussein, Q. M. Hussein, S. A. Abdulwahhab, "Twitter Sentiment Analysis Using an Ensemble Majority Vote Classifier," Journal of Southwest Jiaotong University, 2020: 55: 1-7, 2020.

[34] N. Jiang, F. Tian, J. Li, X. Yuan, J. Zheng, "MAN: mutual attention neural networks model for aspect-level sentiment classification in SIoT," IEEE Internet of Things Journal, 7: 2901-2913, 2020.

[35] U. Naseem, I. Razzak, K. Musial, M. Imran, "Transformer based Deep Intelligent Contextual Embedding for Twitter sentiment analysis," Future Generation Computer Systems, 2020: 1-35, 2020.

[36] M. D. Samad, N. D. Khounviengxay, M. A. Witherow, "Effect of Text Processing Steps on Twitter Sentiment Classification using Word Embedding," arXiv preprint arXiv:2007.13027: 1-14, 2020.

[37] S. M. Nematollahzadeh, S. Ozgoli, M. Sayad Haghighi, "Parameter Identification Method for Opinion Dynamics Models: Tested via Real Experiments," Journal of Electrical and Computer Engineering Innovations (JECEI), 7: 121-131, 2019.

[38] S. Sharma, A. Jain, "An Empirical Evaluation of Correlation Based Feature Selection for Tweet Sentiment Classification," in Proc. Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies, ed: Springer, 2020: 199-208, 2020.

[39] E. Kouloumpis, T. Wilson, J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in Proc. The Fifth International Association for the Advancement of Artificial Intelligence Conf. on Weblogs and Social Media, 2011: 538-541, 2011.

[40] C. D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval 1: Cambridge university press Cambridge, 2008.

[41] S. Chandrakala, C. Sindhu, "Opinion Mining and sentiment classification a survey," ICTACT journal on soft computing, 3: 420-425, 2012.

[42] A. Tripathy, A. Agrawal, S. K. Rath, "Classification of sentiment reviews using n-gram machine learning approach," Expert Systems with Applications, 57: 117-126, 2016.

[43] E. Fersini, E. Messina, F. A. Pozzi, "Sentiment analysis: Bayesian ensemble learning," Decision support systems, 68: 26-38, 2014.

Biographies

Razieh Asgarnezhad received her B.Sc. and MSc. degrees in Computer Engineering from Kashan Azad University in 2009 and Arak Azad University in 2012, respectively. She is currently a Ph.D. candidate at the Department of Computer Engineering at Isfahan Azad University. Her current researches include Data Mining, Text Mining, Learning Automata, and Wireless Sensor Network.

S. Amirhassan Monadjemi is an Associate Professor at the University of Isfahan and Senior Lecturer, School of continuing and lifelong education at the National University of Singapore. He received a Ph.D. in Pattern Recognition from the University of Bristol in 2004. His research interests include Artificial Intelligence, Machine Vision, and Data Analysis.

Mohammadreza SoltanAghaei is an Associate Professor at the Islamic Azad University of Isfahan. He received a Ph.D. in Computer Network from UPM University of Malaysia in 2010. His research interests include Computer Networks and Data Mining.

How to cite this paper:
R. Asgarnezhad, S.A. Monadjem, M. Soltanaghaei, "A High-Performance Model based on Ensembles for Twitter Sentiment Classification, JECEI," Journal of Electrical and Computer Engineering Innovations, 8(1): 41-52, 2020.

DOI: 10.22061/JECEI.2020.7100.357

URL: http://jecei.sru.ac.ir/article_1422.html