Scientific Research Communications

Research Article

Sci. Res. Comm. (2021)
Volume 1, Issue 1
doi: 10.52460/src.2021.004

Ensemble of Classifiers and Term Weighting Schemes for Sentiment Analysis in Turkish

Aytuğ Onan[1,*]

[1] Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Katip Çelebi University, İzmir, Turkey
[*] Corresponding author: aytug.onan@ikcu.edu.tr

Received: 21.03.2021 Accepted: 29.06.2021

Abstract

With the advancement of information and communication technology, social networking and microblogging sites have become a vital source of information. Individuals can express their opinions, grievances, feelings, and attitudes about a variety of topics. Through microblogging platforms, they can express their opinions on current events and products. Sentiment analysis is a significant area of research in natural language processing because it aims to define the orientation of the sentiment contained in source materials. Twitter is one of the most popular microblogging sites on the internet, with millions of users daily publishing over one hundred million text messages (referred to as tweets). Choosing an appropriate term representation scheme for short text messages is critical. Term weighting schemes are critical representation schemes for text documents in the vector space model. We present a comprehensive analysis of Turkish sentiment analysis using nine supervised and unsupervised term weighting schemes in this paper. The predictive efficiency of term weighting schemes is investigated using four supervised learning algorithms (Naive Bayes, support vector machines, the k-nearest neighbor algorithm, and logistic regression) and three ensemble learning methods (AdaBoost, Bagging, and Random Subspace). The empirical evidence suggests that supervised term weighting models can outperform unsupervised term weighting models.

Keywords: Ensemble methods; term weighting schemes; sentiment analysis; text classification

1. Introduction

With the tremendous growth of social media and microblogging sites, the enormous quantity of information available will serve as an important source for decision making, regarding products, services, and policies (Onan, 2017; Onan, 2018). People may express their views, complaints, feelings, and attitudes towards subjects. They can express their ideas about current issues, and products through microblogging platforms.

Twitter is one of the most common microblogging sites in the world, in which millions of people publishing more than one hundred million text messages (referred as, tweets) every single day. On Twitter, users can send short messages with a character limit of 280. On Twitter, users can post messages about real-word events occurring around the world, aside from personal tweets. Many incidents are now being posted on Twitter for the first time, as near real-time as it occurs (Samant et al., 2019). The content created by users on Twitter provides researchers and practitioners with a valuable source of information, which can be employed for several applications, including earthquake prediction (Sakaki et al., 2010), influenza epidemics (Woo et al., 2018), and crisis management (Hecht et al., 2011).

Sentiment analysis is a task in natural language processing, which seeks to identify the semantic orientation of text documents, with the use of tools and techniques from computer science, data science and statistics (Onan and Korukoğlu, 2017). Sentiment analysis can be employed to obtain useful information from unstructured text documents. Traditional application fields for sentiment analysis include the detection of public sentiment for policy-making purposes and the market analysis of goods and services based on feedbacks of consumers (Zhang et al., 2009; Fersini et al., 2014). In that sense, organized, insightful knowledge obtainable by recognizing subjective information from online content can be extremely useful for decision making, including decision support systems and individual decision makers (Onan et al., 2016). The approaches used in sentiment analysis can be divided into two groups, as machine learning-based and lexicon-based methods. In machine learning-based sentiment analysis, the identification of sentiment orientation has been modelled as a text classification problem, in which supervised learners, such as Naïve Bayes, support vector machines and artificial neural networks have been employed (Aggarwal and Zhai, 2012).

For short text messages, the identification of an appropriate term representation scheme is a crucial task. In the vector space model, term weighting schemes are important schemes to represent text documents.

In this paper, we present a comprehensive analysis on sentiment analysis in Turkish with two unsupervised term weighting schemes (i.e., term frequency, and TF-IDF) and seven supervised term weighting schemes (i.e., odds ratio, relevance frequency, balanced distributional concentration, inverse question frequency-question frequency-inverse category frequency, short text weighting, and inverse gravity moment and regularized entropy). Four supervised learning algorithms (i.e., Naïve Bayes, support vector machines, k-nearest neighbor algorithm and logistic regression) and three ensemble learning methods (i.e., AdaBoost, Bagging and Random Subspace) are used to explore the predictive efficiency of the term weighting schemes. To the best of our knowledge, it is the first study in Turkish text sentiment classification, where supervised and unsupervised term weighting schemes have been comprehensively evaluated in conjunction with supervised learning models and ensemble classification models. The experimental results indicate that supervised term weighting models can outperform unsupervised term weighting models.

The remainder of the paper is structured as follows: Related work on sentiment analysis is discussed in Section 2. Section 3 presents the term weighting schemes. Section 4 presents classifiers, Section 5 presents the ensemble methods. The experimental procedure, dataset and the empirical results are discussed in Section 6. The concluding remarks have been presented in Section 7.

2. Related Work

Sentiment analysis on the content created by users on Twitter has attracted great research attention. The early work on sentiment analysis on Twitter data is briefly reviewed in this section.

Go et al. (2009) evaluated the predictive performance of maximum entropy and support vector machine classifiers for sentiment analysis on Twitter messages. In this study, text documents are represented using different structures such as 1-gram, 2-gram, and part-of-speech tags. It was observed that 80% correct classification performance was achieved with the developed method. In another study, Agarwal et al. (2011) examined the effectiveness of 1-gram scheme, feature engineering-based schemes and tree-based data representation for sentiment analysis on Twitter. In the 1-gram representation, the data set is represented by using around 10000 features, while the number of features in the representation based on feature engineering was reduced to 100, but the correct classification performance was kept higher. On the other hand, it has been observed that the highest accuracy classification achievements are obtained when Twitter messages are represented using the tree structure. Similarly, Kouloumpis et al. (2011) evaluated the effectiveness of n-gram features, dictionary-based features, part-of-speech tags, and Twitter-specific features for sentiment analysis on Twitter messages. In the empirical analysis, the highest predictive performance has been obtained by n-gram features. In the study performed by De Boom et al. (2016), word2vec word embedding scheme and a weighted word embedding vector extraction method based on TF-IDF weighting function were presented to capture the semantic integrity on short text documents, such as Twitter messages.

Similarly, Djaballah et al. (2019) evaluates the predictive performance of machine learning and deep learning-based schemes for sentiment analysis on Twitter messages. In experimental analysis, vectors obtained using word2vec word embedding method were considered in two different ways, namely, unweighted, and weighted vector pooling. Experimental results show that weighted vector pooling gives higher performance. The predictive performance of various n-gram models (namely unigram, bigram, and trigram) and their combinations on sentiment analysis of Turkish Twitter messages were examined by Onan (2017). In empirical research, the combination of unigram and bigram features achieves the highest predictive efficiency. In a similar way, Onan (2018) introduced an ensemble approach to sentiment analysis on Twitter based on LIWC (i.e., Linguistic Inquiry and Word Count) categories. In the study performed by Şahin (2017), vectors obtained by word2vec word embedding scheme and support vector machines have been utilized to classify Turkish text documents. In another study, Griol et al. (2020) presented an ensemble classification scheme for sentiment analysis on Twitter messages, which incorporates ensemble of feature sets based on opinion lexicons, n-grams, and word clusters in conjunction with maximum entropy classifier. Similarly, Samant et al. (2019) examined the predictive performance of supervised and unsupervised term weighting schemes for sentiment analysis on Twitter and they presented a novel improved supervised term weighting model. Recently, Carvalho and Plastino (2020) comprehensively examined the predictive performance of n-gram features, meta-level features, microblog features, part-of-speech features, surface features, emoticon features, and word embedding based features on sentiment analysis for Twitter messages.

3. Term Weighting Schemes

Term weighting schemes can be classified predominantly into two categories, as unsupervised and supervised term weighting schemes (Samant et al., 2019). For unsupervised term weighting schemes, category information is not utilized to allocate weight values to words, whereas supervised term weighting schemes use category information from the training data for a particular term. Let N denote the total number of documents in the corpus, let tf denote the frequency of the word indicating the number of times in the document a specific term has been encountered, and let df denote the number of documents in which at least one specific term has been encountered.

Term frequency (*tf*) is an unsupervised term weighting scheme to compute weight value w_{d_i,t_j} for term t_j in document d_i, as given by Equations 1 and 2 (Samant et al., 2019):

$$w_{d_i,t_j} = tf \tag{1}$$

$$w_{d_i,t_j} = \begin{cases} 1 & , \quad term\ encountered \\ 0 & , \quad term\ not\ encountered \end{cases} \tag{2}$$

Term frequency-inverse document frequency (TF-IDF) is another unsupervised term weighting scheme on information retrieval and text mining. Term frequency represents the relative frequency of a word *t* in a text document and inverse document frequency scales with the number of documents. *TF-IDF* weighting scheme can be computed as given by Equation 3:

$$w_{d_i,t_j} = tf * log\left(\frac{N}{df}\right) \tag{3}$$

Odds ratio (*OR*) is a supervised term weighting scheme, which is the ratio of the probability of occurrence of an event in one group to the probability of occurrence in another group. *OR* can be computed as given by Equation 4 (Quan et al., 2010):

$$OR = log\left(\frac{tp * tn}{fp * fn}\right) \tag{4}$$

where *tp* denotes true positives, *tn* denotes true negatives, *fp* denotes false positives and *fn* denotes false negatives.

Relevance frequency (*RF*) is another supervised term weighting scheme. In this scheme, the ratio of number of positive category (positive class label) documents consisting of the word to the number of negative category (negative class label) documents containing the word has been considered to compute weight values, as given by Equation 5 (Lan et al., 2006):

$$RF = log\left(2 + \frac{tp}{Max(1, fn)}\right) \tag{5}$$

Balanced distributional concentration (*bdc*) is another supervised term weighting scheme based on entropy. Balanced distributional concentration test term *t*'s discriminating power based on its distribution in different categories (*c_i*). Balanced distributional concentration can be computed, as given by Equation 6 (Wang et al., 2015):

$$bdc = 1 - \frac{BH_t}{\log{(K)}} \tag{6}$$

$$BH_t = -\sum_{i=1}^{K} \frac{p(t/c_i)}{\sum_{i=1}^{K} p(t/c_i)} \log\left(\frac{p(t/c_i)}{\sum_{i=1}^{K} p(t/c_i)}\right) \tag{7}$$

where K denotes total number of categories in the training data and $p(t/c_i)$ denotes the probability of term t in category c_i.

Inverse question frequency-question frequency-inverse category frequency (IQF) is another supervised term weighting scheme for short text classification, which can be computed as given by Equation 8 (Quan et al., 2010):

$$iqf = log\left(\frac{N}{tp + fn}\right) * \log(tp + 1) * log\left(\frac{K}{cf} + 1\right) \tag{8}$$

where *cf* denotes the number of categories that have at least one document in which *t* has been encountered. Short text weighting (SW) is another supervised term weighting scheme for short text classification, which can be calculated as given by Equation 9:

$$W(t_{ij}) = \frac{tf_{ij} + 1}{\sum_{j=1}^{|T|} tf_{ij} + |T|} * log\left(1 + \frac{tp}{fp + fn + 1}\right) \tag{9}$$

where j denotes term present in document i, which contains $|T|$ terms and tf_{ij} denotes the frequency of it.

Inverse gravity moment (*IGM*) is another supervised term weighting scheme based on class-specific gravity (Chen et al., 2016). Inverse gravity moment-based weight value for a term t_i has been computed as given by Equations 10 and 11:

$$W(t_{ij}) = 1 + \lambda * IGM(t_i) \tag{10}$$

$$IGM(t_i) = \frac{f_{i1}}{\sum_{r=1}^{K} f_{ir} * r} \tag{11}$$

where λ is the parameter which has been set according to (Samant et al., 2019) and f_{ir} denotes term's frequency in category *r.*

Regularized entropy (*RE*) is another supervised term weighting scheme, which seeks to find a balanced weighting scheme for terms by measuring term distribution. Regularized entropy can be computed as given by Equation 12 (Wu et al., 2017):

$$RE = b + (1 - b) * (1 - h), where\ b \in [0,1] \tag{12}$$

$$h = -p^{+} * \log(p^{+}) - p^{-} * \log(p^{-}) \tag{13}$$

$$p^{+} = -\frac{tp/(tp + fp)}{\frac{tp}{tp + fp} + \frac{tp}{fn + tn}} \tag{14}$$

$$p^{-} = -\frac{fn/(fn + tn)}{\frac{tp}{tp + fp} + \frac{tp}{fn + tn}} \tag{15}$$

4. Supervised Learning Models

Naïve Bayes, support vector machines, k-nearest neighbour, and logistic regression algorithm are used to evaluate the predictive efficiency of unsupervised and supervised term weighting schemes.

Naïve Bayes algorithm (NB) is a probabilistic classification algorithm based on Bayes' theorem. Owing to the assumption of conditional independence, it has a basic structure. It can be used efficiently in text and web mining applications, despite its simple structure (Onan, 2016).

Support vector (SVM) machines are supervised learning algorithms that can be used to solve problems with classification and regression based on maximum margin hyperplane. To classify both linear and non-linear data, they can be applied effectively (Onan, 2017). To solve classification or regression problems, support vector machines construct a hyperplane in a higher dimensional space. By reaching the greatest distance to the nearest training data points of classes, the hyperplane seeks to make a good separation.

An instance-based classifier is the K-nearest Neighbour Algorithm (KNN). The class label of each instance in the KNN algorithm is calculated based on the k-nearest neighbours of the instance. A majority voting mechanism is used to decide the class label, based on the predictions of neighbouring instances.

Logistic regression (LR) is a linear classification algorithm that uses a linear function of a collection of predictor variables to model the likelihood of occurrence of any event (Kantardzic, 2011). Linear regression can provide good outcomes. The membership values produced by linear regression, however, cannot always be within the [0-1] range, which is not an acceptable probability range. A linear model is based on the transformed target variable in logistic regression, while removing the stated problem.

5. Ensemble Learning Models

Three ensemble learning methods (i.e., AdaBoost, Bagging and Random Subspace) are used to explore the predictive efficiency of the term weighting schemes.

The AdaBoost algorithm is a common method of ensemble learning that aims to obtain a robust classification system by focusing on hard-to-classify data points (Freund and Schapire, 1996). The weight values assigned to the training set instances are modified in this method such that the weight values of misclassified instances are increased, while the weight values of properly classified instances are reduced. The learning algorithms therefore concentrate on the classification of difficult cases.

Bagging (Bootstrap aggregating) is a common method of ensemble learning that aims to achieve a single prediction with higher predictive output by combining weak algorithms of learning trained on different training sets (Breiman, 1996). Different training sets are obtained in this scheme by simple random sampling with substitution. By majority voting or weighted voting, the forecasts of poor learning algorithms are combined.

The random subspace algorithm is an ensemble learning algorithm that combines many classifiers trained on randomly selected subspaces of functions (Ho, 1998). The algorithm aims to avoid over-fitting by training the weak learning algorithms on various samples of the feature space, thus providing high predictive efficiency.

6. Experimental Procedure and Empirical Results

This section presents the experimental procedure, evaluation measures, and the empirical results of the study.

In the experimental analysis, a data set containing Turkish Twitter messages was created to evaluate the performance of the unsupervised and supervised term weighting schemes. The dataset was acquired over a two-month period using an application written in Python using the Twitter API. In the sentiment analysis dataset, there are a total of 21000 Twitter messages, 10500 of which are positive and 10500 are negative. On the data set, pre-processing steps such, as stemming, extraction of stop words, and root finding were applied. To annotate raw Twitter messages, we used an annotation process in which each message was assigned to one of two categories based on its sentiment orientation: positive or negative. The raw messages have been annotated by two experts. Cohen's kappa (κ) metric has been calculated. We obtained a score of 0.82 for the corpus, indicating perfect agreement between the annotators.

During the stemming phase, Twitter messages reporting both a positive and a negative statement were removed from the data set. In addition, each of the letters in the messages has been converted to lowercase letters, punctuation marks, numbers, and special characters such as '@', '#' have been removed. Text messages were filtered by terms and character length, duplicate letters were removed. Lucene application development interface was used to extract the stop words, and Zemberek library was used in the root finding phase.

The performance of the data set represented by these data representation methods was evaluated with four basic classifiers, namely, k-nearest neighbour algorithm (KNN), support vector machines (SVM), logistic regression (LR) and Naive Bayes (NB). In addition, three ensemble learning methods (i.e., AdaBoost, Bagging and Random Subspace) are used to explore the predictive efficiency of the term weighting schemes. In machine learning based experimental analysis, 10-fold cross validation was used. Implementation was done with WEKA 3.9 via default parameter values.

To evaluate the performance of term weighting schemes, classification algorithms and ensemble learning methods, classification accuracy and F-measure have been utilized.

Classification accuracy (ACC) is the proportion of true positives and true negatives obtained by the classification algorithm over the total number of instances as given by Equation 16:

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \tag{16}$$

where *TN* denotes number of true negatives, *TP* denotes number of true positives, *FP* denotes number of false positives and *FN* denotes number of false negatives.

Precision (PRE) is the proportion of the true positives against the true positives and false positives as given by Equation 17:

$$PRE = \frac{TP}{TP + FP} \tag{17}$$

Recall (REC) is the proportion of the true positives against the true positives and false negatives as given by Equation 18:

$$REC = \frac{TP}{TP + FN} \tag{18}$$

F-measure takes values between 0 and 1. It is the harmonic mean of precision and recall as determined by Equation 19:

$$F - measure = \frac{2 * PRE * REC}{PRE + REC} \tag{19}$$

In Tables 1 and 2, the classification accuracy values, and F-measure values obtained by unsupervised and supervised term weighting schemes on conventional classification algorithms and ensemble learning methods have been presented, respectively. As it can be observed from the empirical results listed in Table 1, supervised term weighting schemes outperform the unsupervised term weighting schemes for short text classification in Turkish.

Regarding the empirical results presented in Tables 1 and 2, the highest predictive performance among the conventional classification algorithms has been generally obtained Naïve Bayes algorithm in conjunction with supervised and unsupervised term weighting schemes. Ensemble learning models outperform the conventional classification schemes when term weighting-based text representation models have been utilized for text representation in Turkish. The highest predictive performances among all the compared schemes have been generally obtained by random subspace ensemble of support vector machines. Among all the compared supervised and unsupervised term weighting schemes, regularized entropy (RE) outperforms the other term weighting schemes. The highest predictive performance among all the compared configurations has been achieved by regularized entropy-based term weighting in conjunction with random subspace ensemble of support vector machines.

Table 1. Classification accuracies by learning algorithms and term weighting methods

Weighting Scheme	TF	TF-IDF	OR	RF	BDC	IQF	SW	IGM	RE
NB	79.60	78.15	82.72	82.92	83.19	83.37	82.98	83.24	83.31
SVM	78.88	77.43	82.04	82.33	82.55	82.75	82.52	82.74	82.79
KNN	72.38	70.93	77.20	77.68	78.04	78.24	77.87	78.19	78.32
LR	76.53	75.08	79.88	80.19	80.46	80.65	80.31	80.48	80.94
AdaBoost (NB)	82.70	81.25	86.26	86.42	86.83	87.11	87.22	87.61	88.19
AdaBoost (SVM)	80.37	78.92	83.67	83.91	84.17	84.35	84.02	84.28	84.39
AdaBoost (KNN)	79.39	77.94	82.63	82.82	83.09	83.41	83.05	83.22	83.40
AdaBoost (LR)	73.82	72.37	78.58	79.05	79.37	79.48	79.17	79.49	79.61
Bagging (NB)	76.64	75.19	79.96	80.23	80.58	80.67	80.37	80.89	80.97
Bagging (SVM)	78.72	77.27	81.87	82.08	82.34	82.54	82.14	82.38	82.45
Bagging (KNN)	79.39	77.94	82.63	82.85	83.11	83.42	83.09	83.33	83.45
Bagging (LR)	80.37	78.92	83.81	83.93	84.27	84.44	84.13	84.33	84.39
Random Subspace (NB)	81.51	80.06	84.65	84.78	85.06	85.28	84.94	85.40	85.58
Random Subspace (SVM)	85.58	84.13	88.88	88.99	89.27	89.34	89.17	89.42	89.76
Random Subspace (KNN)	85.13	83.68	88.28	88.57	88.94	88.99	88.83	88.97	89.04
Random Subspace (LR)	82.48	81.03	86.21	86.41	86.80	86.93	86.67	87.48	87.77

Table 2. F-measure values obtained by learning algorithms and term weighting methods

Weighting Scheme	TF	TF-IDF	OR	RF	BDC	IQF	SW	IGM	RE
NB	0.80	0.79	0.83	0.84	0.84	0.84	0.84	0.84	0.84
SVM	0.80	0.78	0.83	0.83	0.83	0.83	0.83	0.83	0.83
KNN	0.73	0.72	0.78	0.78	0.79	0.79	0.78	0.79	0.79
LR	0.77	0.76	0.81	0.81	0.81	0.81	0.81	0.81	0.82
AdaBoost (NB)	0.83	0.82	0.87	0.87	0.88	0.88	0.88	0.88	0.89
AdaBoost (SVM)	0.81	0.80	0.84	0.85	0.85	0.85	0.85	0.85	0.85
AdaBoost (KNN)	0.80	0.79	0.83	0.83	0.84	0.84	0.84	0.84	0.84
AdaBoost (LR)	0.74	0.73	0.79	0.80	0.80	0.80	0.80	0.80	0.80
Bagging (NB)	0.77	0.76	0.81	0.81	0.81	0.81	0.81	0.82	0.82
Bagging (SVM)	0.79	0.78	0.83	0.83	0.83	0.83	0.83	0.83	0.83
Bagging (KNN)	0.80	0.79	0.83	0.84	0.84	0.84	0.84	0.84	0.84
Bagging (LR)	0.81	0.80	0.84	0.85	0.85	0.85	0.85	0.85	0.85
Random Subspace (NB)	0.82	0.81	0.85	0.85	0.86	0.86	0.86	0.86	0.86
Random Subspace (SVM)	0.86	0.85	0.90	0.90	0.90	0.90	0.90	0.90	0.90
Random Subspace (KNN)	0.86	0.84	0.89	0.89	0.90	0.90	0.90	0.90	0.90
Random Subspace (LR)	0.83	0.82	0.87	0.87	0.88	0.88	0.87	0.88	0.88

We observed that the relative performance of term weighting schemes varies significantly across datasets and classifiers in our experiments. Unsupervised schemes outperformed supervised schemes, while regularized entropy-based schemes outperformed supervised schemes. The empirical results indicate that the weight values for terms can be improved by employing supervised term weighting models which utilize class-specific information.

To summarize the main findings of the empirical results, Figure 1 denotes the bar chart for classification accuracies based on supervised learning models and ensemble learning methods and Figure 2 illustrate the bar chart for classification accuracies based on term weighting schemes.

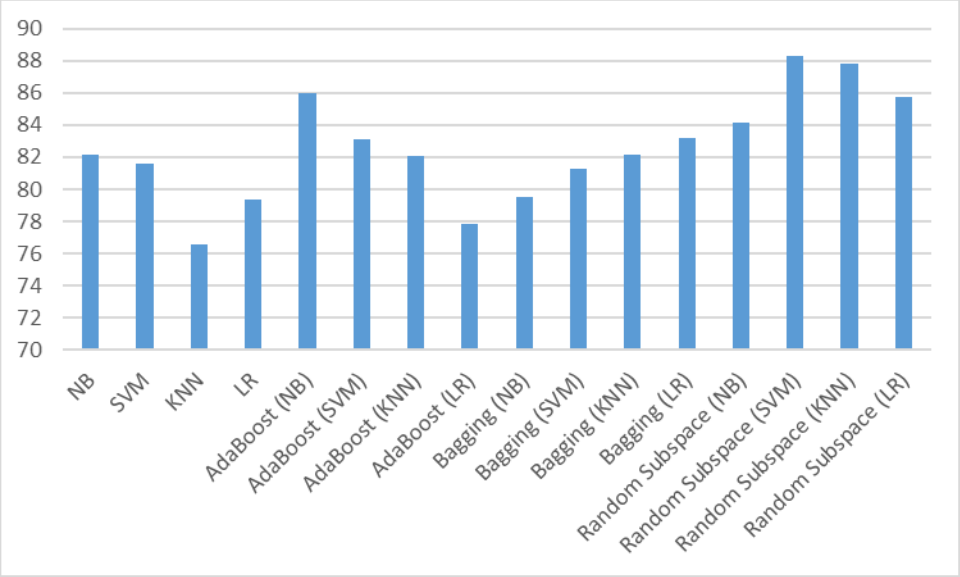

Figure 1. The bar chart for accuracy for classifiers

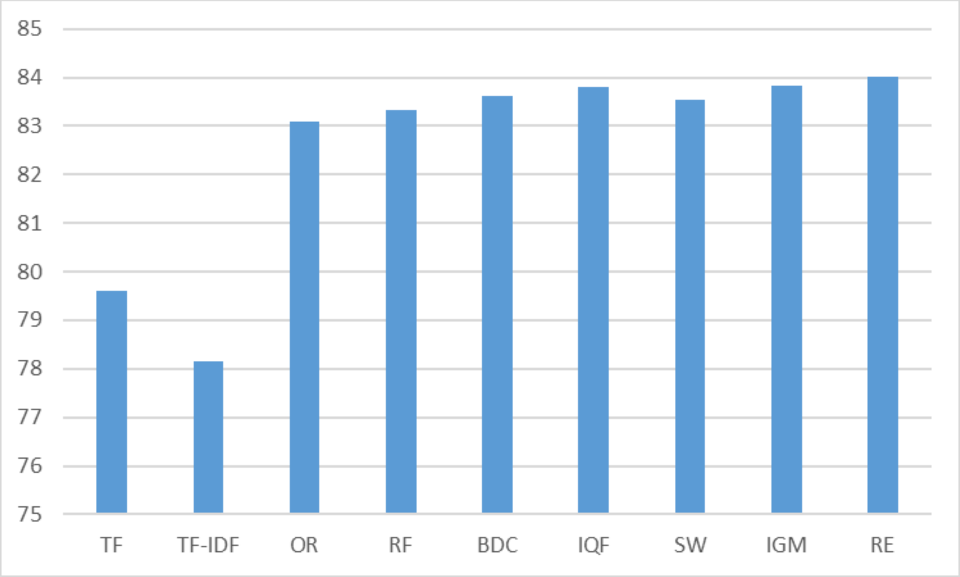

Figure 2. The bar chart for accuracy for weighting schemes

7. Conclusions

Social networking and microblogging sites have developed into important sources of information as information technology advances. Individuals can express their opinions, concerns, thoughts, and attitudes on a wide variety of subjects. They should make use of microblogging platforms to voice their opinions about current events and products. Sentiment analysis is a critical subfield of natural language processing research that aims to characterize the sentiment orientation of source materials. Twitter is one of the world's most popular microblogging platforms, with millions of users posting over a hundred million text messages daily (known as tweets). The task of choosing an appropriate scheme for representing terms in short text messages is critical. Term weighting schemes are advantageous when it comes to representing text documents in a vector space model. We present a comprehensive analysis of Turkish sentiment analysis in this paper, utilizing nine supervised and unsupervised term weighting schemes. To investigate the predictive efficiency of term weighting schemes, four supervised learning algorithms (Naive Bayes, support vector machines, k-nearest neighbor algorithm, and logistic regression) and three ensemble learning methods (AdaBoost, Bagging, and Random Subspace) are used. The results of the experiments indicate that supervised term weighting models can outperform unsupervised term weighting models. Regularized entropy (RE) outperforms all other term weighting schemes when supervised and unsupervised schemes are compared. Regularized entropy-based term weighting in combination with a random subspace ensemble of support vector machines produced the best predictive performance of all the configurations compared.

Author Statement

The author confirms sole responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation.

Conflict of Interest

The author declares no conflict of interest.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30-38).

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163-222). Springer, Boston, MA.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Carvalho, J., & Plastino, A. (2020). On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis. *Artificial Intelligence Review*, 1-50.

De Boom, C., Van Canneyt, S., Demeester, T., & Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, *80*, 150-156.

Djaballah, K. A., Boukhalfa, K., & Boussaid, O. (2019, October). Sentiment Analysis of Twitter Messages using Word2vec by Weighted Average. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 223-228). IEEE.

Fersini, E., Messina, E., & Pozzi, F. A. (2014). Sentiment analysis: Bayesian ensemble learning. *Decision support systems*, *68*, 26-38.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In *icml* (Vol. 96, pp. 148-156).

Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009.

Griol, D., Kanagal-Balakrishna, C., & Callejas, Z. (2020). Feature Set Ensembles for Sentiment Analysis of Tweets. In *Advances in Data Science: Methodologies and Applications* (pp. 189-208). Springer, Cham.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011, May). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 237-246).

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832-844.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Kouloumpis, E., Wilson, T., & Moore, J. (2011, July). Twitter sentiment analysis: The good the bad and the omg!. In *Fifth International AAAI conference on weblogs and social media*.

Lan, M., Tan, C. L., & Low, H. B. (2006, July). Proposing a new term weighting scheme for text categorization. In *AAAI* (Vol. 6, pp. 763-768).

Onan, A. (2016). Classifier and feature set ensembles for web page classification. *Journal of Information Science*, *42*(2), 150-165.

Onan, A. (2017). Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı duygu analizi. *Yönetim Bilişim Sistemleri Dergisi*, *3*(2), 1-14.

Onan, A. (2017, April). Sarcasm identification on twitter: a machine learning approach. In *Computer Science On-line Conference* (pp. 374-383). Springer, Cham.

Onan, A. (2018). Sentiment analysis on Twitter based on ensemble of psychological and linguistic feature sets. *Balkan Journal of Electrical and Computer Engineering*, *6*(2), 69-77.

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, *62*, 1-16.

Quan, X., Wenyin, L., & Qiu, B. (2010). Term weighting schemes for question categorization. *IEEE transactions on pattern analysis and machine intelligence*, *33*(5), 1009-1021.

Şahin, G. (2017, May). Turkish document classification based on Word2Vec and SVM classifier. In *2017 25th signal processing and communications applications conference (SIU)* (pp. 1-4). IEEE.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860).

Samant, S. S., Murthy, N. B., & Malapati, A. (2019). Improving Term Weighting Schemes for Short Text Classification in Vector Space Model. *IEEE Access*, *7*, 166578-166592.

Wang, T., Cai, Y., Leung, H. F., Cai, Z., & Min, H. (2015, November). Entropy-based term weighting schemes for text categorization in VSM. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 325-332). IEEE.

Woo, H., Cho, H. S., Shim, E., Lee, J. K., Lee, K., Song, G., & Cho, Y. (2018). Identification of keywords from twitter and web blog posts to detect influenza epidemics in Korea. *Disaster medicine and public health preparedness*, *12*(3), 352-359.

Wu, H., Gu, X., & Gu, Y. (2017). Balancing between over-weighting and under-weighting in supervised term weighting. *Information Processing & Management*, 53(2), 547-557.

Zhang, C., Zeng, D., Li, J., Wang, F. Y., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology*, *60*(12), 2474-2487.