ITC 3/47

Journal of Information Technology and Control
Vol. 47 / No. 3 / 2018
pp. 457-470
DOI 10.5755/j01.itc.47.3.20935
© Kaunas University of Technology

Extended Feature Spaces Based Classifier Ensembles for Sentiment Analysis of Short Texts

Received 2018/06/12 | Accepted after revision 2018/08/16

http://dx.doi.org/10.5755/j01.itc.47.3.20935

Extended Feature Spaces Based Classifier Ensembles for Sentiment Analysis of Short Texts

Zeynep Hilal Kilimci

Faculty of Engineering; Dogus University; Acıbadem, Kadıköy, 34722, İstanbul, Turkey; phone: +90 216 444 7997; fax: +90 216 327 9631; e-mail: hkilimci@dogus.edu.tr

Sevinc Ilhan Omurca

Faculty of Engineering; Kocaeli University; Umuttepe Yerleşkesi, 41380, Kocaeli, Turkey; phone: +90 262 303 3572; fax: +90 262 303 1033; e-mail: silhan@kocaeli.edu.tr

Corresponding author: hkilimci@dogus.edu.tr

Sentiment classification has become very popular to analyze opinions about events, products, and so on, especially for social networks such as Twitter. Due to the size limitation of expressing ideas on social networks, the classification performance needs to be boosted by proposing various techniques. In this work, the enhancement of feature space with word embedding based features is proposed to deal with the size limitation issues and the classification success of sentiment analysis is improved by employing classifier ensembles. The contributions of this paper are fivefold. First, the representative capabilities of features are enriched by using a semantic word embedding model and followingly the conventional feature selection techniques are compared. Second, traditional machine learning algorithms, namely naïve Bayes, support vector machine, and random forest are carried out to select baseline classifier for the proposed ensemble system. Third, three ensemble strategies namely, bagging, boosting, and random subspace are introduced to ensure the diversity of ensemble learning. Fourth, experiments are conducted to compare the performance of the models with the word embedding baseline. Eventually, a wide range of comparative experiments on Twitter datasets demonstrate that the classification performance of the proposed model significantly outperforms the state-of-the-art studies.

KEYWORDS: Word embedding, ant colony optimization, information gain, sentiment analysis, classifier ensembles, extended spaces.

1. Introduction

Social media has become a very popular resource to analyze huge amount of information and detect opinions on many things about various subjects on the Internet. As one of the well-known social media platforms, Twitter is preferred by up to 100 million active users to express opinions. This means that Twitter comprises precious information which can be effective for market dynamics. For this reason, the sentiment analysis is a significant part to understand user demands in terms of positive and negative aspects.

Sentiment analysis is a considerable research field and can be summarized as the extraction of users' opinions from the text. The traditional machine learning techniques such as naïve Bayes, support vector machines, and so on are employed to determine the sentiment polarity such as negative, positive, or neutral on this domain. The most popular and recently used one is deep learning models used to achieve higher classification performance compared to the conventional machine learning algorithms. The fundamental approach of deep learning models is to provide automatic feature extraction by training complex features with minimum external support and acquire the meaningful representation of data through deep neural networks for sentiment analysis. For this purpose, many networks such as convolutional neural networks (CNN), recurrent neural networks (RNN), recursive neural networks, deep belief networks (DBN), and various semantic word embedding models such as word2vec, Glove are employed. These techniques have been extensively applied by researchers in different areas such as computer vision, image analysis, speech recognition, and natural language processing.

As much as the selection of classifier, the individual success and diversity of base learners are also determinative factors of the ensemble success. As the diversity of base learner increases, the classification success of system becomes better. The usage of different or the same base learners is requisite in order to provide diversity. Diversity is maintained with several conventional ensemble algorithms such as bagging, random subspaces, random forests, and rotation forest for the same base learners. For different base learners, it is already achieved by blending different learning algorithms with various decision making techniques such as majority voting, stacking, cascading. In this work, we focus on the homogeneous classifier ensembles which utilize the same base learners to provide diversity.

This paper proposes to integrate word embedding approach and ensemble learning models to boost classification performance of short texts by extending feature space. In this study, we centered on enhancing feature space to advance the classification success of short texts because of the size limitation of expressing ideas on social networks such as Twitter. In particular, this work considers an ensemble of classifiers, where classifiers are trained with extended feature spaces by making use of word embedding based feature extraction technique, namely word2vec. The advantage of word embedding based feature extraction methods is to employ semantic word embeddings, on the contrary, traditional feature selection techniques ignore semantically similar words. Followingly, three ensemble strategies namely, bagging, boosting, and random subspace are carried out to ensure the diversity of ensemble learning by choosing the best classification performance of baseline classifier among multinomial naïve Bayes (MNB), multivariate naïve Bayes (MVNB), support vector machine (SVM), and random forest (RF) algorithms. To the best of our knowledge, this is the very first approach of utilizing word embedding based extended spaces with classifier ensembles for short sentiment classification on Twitter. For demonstrating the contribution of proposed model, we conduct experiments on Twitter datasets. Extensive experiments show that the word embedding based proposed model is highly efficient for sentiment analysis compared to the traditional ensemble models.

The rest of the paper is organized as follows: Section 2 gives related researches on the use of deep learning models and word embeddings, sentiment analysis and ensemble systems. In Section 3, the proposed framework is represented. Experiment setup and results are demonstrated in Sections 4 and 5. Section 6 concludes the paper with a discussion and conclusions.

2. Related Work

Many researchers focus on deep learning approach to ensure more accurate classification models for sentiment analysis. Liao et al. [21] propose to comprehend the sentiment analysis of Twitter data employing deep learning models. They compose a simple convolutional neural network model and present better classification performances compared to the traditional learning algorithms such as SVM and naïve Bayes classifiers. A novel deep convolutional neural network which employs from character to sentence level knowledge to carry out sentiment analysis on short texts is recommended by Santos and Gatti [31]. They report that their approach outperforms results of state-of-the-art studies and achieves sentiment classification accuracy with 86.4% on STS corpus. Another work [17] emphasizes the significance of keywords to interpret the semantics. Long short memory and gated recurrent unit are carried out on IMDB and SemEval-2016 datasets by establishing keyword vocabulary. Experiment results show that the efficiency of proposed model of them is verified with 1%-2% accuracy improvement. Sentiment classification of Chinese micro-blogs becomes focus of attention by utilizing improved recurrent neural network model in [11]. They find a way out to solve a long-term dependency by substituting the hidden layer of recurrent neural network with long short term memory structure. Classification success of the system outperforms conventional machine learning algorithm namely, support vector machine with 3.17% precision rate. Another study [39] on sentiment classification aims at employing a new recurrent random walk network by making use of posted tweets and social relations, named as heterogeneous microblog sentiment classification (MSC). The proposed model is based on deep neural networks with random-walk layer by performing the back-propagation method on the training phase. Experiments are carried out on the well-known and widely used datasets from Twitter to demonstrate the success of their model. The proposed technique exhibits better classification performance than other state-of-the-art studies. An efficient translation free deep neural network architecture is adverted in [6] to implement multilingual sentiment analysis on Twitter dataset. The significant part of the proposed model is based on word and character level embeddings by using long short term memory and convolutional networks, respectively. They compare character based architecture with long short term memory embedding, convolutional embedding, convolutional embedding freeze, convolutional character level embedding, and conventional support vector machine algorithm in terms of accuracy and f1-score as evaluation metrics. Extensive experiment results represent that the proposed technique (convolutional character based architecture) is efficient for multilingual sentiment analysis compared to the state-of-the-art deep neural models. In [35], Uysal and Murphey concentrate on the comparison of conventional feature selection models and deep learning approaches for document level sentiment classification. Two types of feature extraction models are exploited in this comparative work. First one is based on term frequency without taking into account order of terms in the document while second is grounded on the term dependencies by making use of semantic word embedding. SVM classifier with linear kernel is utilized to demonstrate the classification performance of traditional approaches. Furthermore, the authors evaluate deep learning based approaches for classification task in this study although these are generally used for the feature selection step on sentiment classification. They report that the proposed deep learning based models with one-hot vectors or fine-tuned semantic word embeddings achieve better results than the word embedding without tuning technique.

There are also several studies on classifier ensembles with extended space. The influential study by Amasyalı and Ersoy [3] proposes the extended feature space by choosing new features randomly and adding them to original feature space. They observe that all extended versions outperform original versions for all ensemble algorithms. To get higher classification performance of ensemble system, they suggest utilizing the extended space methods. The recent studies [1-2] on extended space decision trees propose to increase the ensemble accuracy by suggesting another approach. Instead of randomly producing, new features with high classification capacity are generated by computing the gain ratio of each different candidate features. Thus, they combine newly generated features and existing features in order to extend feature space. The authors conclude that the extended space forest, which means the usage of one more than decision trees, is an effective method to increase prediction accuracy but it can be improved by using significant features instead of selecting randomly.

There are limited studies on the combination of ensemble strategies and word embedding methodology for sentiment classification task. The proposed multilayer perceptron based ensemble model is utilized for predicting sentiment score of financial texts as optimistic or pessimistic in [14]. For this purpose, the authors use four models namely, CNN, LSTM, vector averaging and feature driven to obtain diversity of feature vector by composing a new feature vector at the feature ensembling step. After implementing ensembling step, multilayer perceptron network is utilized as a classifier. Experimental results show that the performance of ensemble of deep learning and feature based models represents remarkable results. Nozza et al. [26] propose to address the problem of domain adaptation by evaluating deep learning and ensemble techniques for sentiment classification. Naïve Bayes, support vector machine, voted perceptron, decision tree, logistic regression, k-nearest neighbour, and random forest are considered as base learners. Bagging, boosting, random subspace, and simple voting are utilized as ensemble methods meantime deep learning part is composed of the autoencoder which is a particular class of artificial neural network. The authors conclude the study claiming that accuracy results of the proposed approach demonstrate considerable enhancement compared to the state-of-the-art studies. Another recent work [4] on deep learning sentiment analysis with ensemble techniques proposes to enhance the success of deep learning techniques by combining them with conventional surface models. For this objective, they focus on deep learning based classifier using a word embeddings model and a linear machine learning algorithm which is employed as a base learner of the ensemble system. Then, ensemble strategy is implemented to combine base learner and other surface classifiers. Extensive comparative experiments demonstrate that the success of proposed techniques outperforms original versions in terms of F1-score.

In this work, the enhancement of feature space with word embedding based features is proposed to deal with the size limitation issues and the classification success of sentiment analysis is improved by employing classifier ensembles. Our work differs from the above mentioned studies in that this is the very first attempt of using word embedding based extended spaces with classifier ensembles on the short-text sentiment classification. The details of the proposed study can be found in Section 3.

3. Proposed Framework

This section introduces our proposed system for the short-text sentiment classification. First, word embeddings and traditional feature selection methods are introduced for the extended feature spaces. After that, the proposed word embedding based model with ensemble strategy is represented.

3.1. Word Embedding (WE)

As noted in the previous works [1-3] the enrichment of feature space ensures significant contribution to the classification performance on the numeric data. The studies so far on extended space forests utilize either randomly chosen features [3] or the specific feature selection method such as gain ratio [1-2] to determine new candidate features to be consolidated to the original feature space. In this study, word embeddings are utilized for the first time to extend original feature space with classifier ensembles using word2vec tool instead of conventional feature selection techniques.

Word2vec is a tool that is used to generate word embeddings by using a group of models. These models propose to reconstruct linguistic contexts of words by employing trained two-layer neural networks. In other words, word embedding tries to discover better word representations of words in a document collection (corpus). The idea behind all of the word embedding is to capture as much contextual, semantical, and syntactical information as possible from documents from a corpus. Word embedding is a distributed representation of words where each word is represented as real-valued vector in a predefined vector space. Distributed representation is based on the notion of distributional hypothesis in which words with similar meaning occur in similar contexts or textual vicinity. Distributed vector representation has proven to be useful in many natural language processing applications such as named entity recognition, word sense disambiguation, machine translation, and parsing [38].

Word2vec is based on two model architectures namely, continuous bag-of-words (CBOW) and continuous skip-gram to perform a distributed representation of words. CBOW model predicts a word given its surrounding context words by ignoring the order of context like bag-of-words approach. On the other hand, continuous skip-gram model aims to predict surrounding context words given a word. In this work, we focus on the continuous skip-gram model due to its considerable performance for infrequent words compared to the CBOW model.

3.2. Information Gain (IG)

The information gain evaluates the number of bits of information obtained for class prediction by knowing the occurrence or nonoccurrence of a feature [10, 34, 40]. In other words, the set of the most significant features with high classification success is acquired for adding to the original feature space. Indeed, the overall feature selection process is to count for score each feature in accordance with a certain feature selection method, and then pick up the best k features.

$$\begin{aligned} IG(t) &= \textstyle\sum_{i=1}^{C} P(C_i)logP(C_i) + \\ &P(t)\textstyle\sum_{i=1}^{C} P(C_i|t)logP(C_i|t) + \\ &P(t')\textstyle\sum_{i=1}^{C} P(C_i|t')logP(C_i|t'), \end{aligned} \tag{1}$$

where C represents the number of classes and $P(C_i)$ demonstrates the probability of C_i, $P(t)$ and $P(t')$ symbolizes the probability of presence and absence of term t', respectively.

3.3. Ant Colony Optimization (ACO)

The ant colony optimization is an optimization technique that can be also employed for feature selection on various domains. It is based on finding the shortest paths from the nest to food source by means of pheromone trails, which is an odorous substance and is excreted by ants. Therefore, the deposition of pheromone is the fundamental factor in order to discover the shortest paths over a certain period of time. Ants mark the path from the nest to a source of food by means of pheromone once they discover a source of food. Then, each isolated ant acts by following direction rich in this substance. That is, the way excreted pheromone is used by more ants and pheromone trails probabilistically enforce to choose the previously marked path for each isolated ant. On less preferred paths, pheromone evaporates over time and the shortest path is discovered by means of the higher ratio of ant traversals. For this reason, there is a transition probabilistic rule for each ant to determine the probability of being selected corresponding path. Hence, ant colony optimization (ACO) technique is attractive for feature selection process that can direct search to optimal subset every time. The probabilistic transition rule, expressing the probability of an ant at feature i choosing to travel to feature j at time t, is as follows:

$$p_{ij}^{k}(t) = \left(\begin{array}{ll} \sum_{l\in J_i^k} \frac{[\tau_{ij}(t)^{\alpha}][\eta_{ij}{}^{\beta}]}{[\tau_{il}(t)^{\alpha}][\eta_{il}{}^{\beta}]} & \text{if } j\in J_i^k, \\ 0 & \text{otherwise} \end{array}\right), \tag{2}$$

where k is the number of ants, η_{ij} is the heuristic desirability of selecting feature j when at feature i, $J^k{}_i$ is the set of ant k's unvisited features, and $\tau_{ij}(t)$ is the amount of virtual pheromone on edge (i,j), α provides global information and determines the relative importance of the pheromone value, β is the heuristic information and presents local information. Producing a number of k ants is the first step for ACO feature selection process. In this study, the number of ants is set equal to the number of features within dataset. Thus, each ant begins with one random feature and they travel edges probabilistically until stopping gauge is fulfilled. The subsets are congregated and then evaluated. Once the algorithm has performed a certain number of times or an optimal subset has attained, the overall feature selection process terminates by obtaining the best feature output. If neither condition holds, it is inevitable to update the intensity of pheromone, then new ants are produced and the feature selection process reiterates once more. The pheromone update is realized by the following rule on each edge:

$$\tau_{ij}(t+1) = (1-\rho)\,\tau_{ij}(t) + \rho\Delta\tau_{ij}(t), \tag{3}$$

where ρ is the pheromone evaporation/update coefficient, $\Delta\tau_{ij}(t)$ denotes quantity of pheromone deposited by each ant k.

3.4. Extended Feature Space

After obtaining semantically the most significant words and word embeddings with the techniques mentioned above, the next step will be to enrich the feature space with these methods. Ultimately, three types of extended feature space are obtained and the first two are constituted with the traditional feature selection techniques. The first extended feature space comprises the combination of original features and significant ones picked up with information gain technique (original + IG). The second feature space is enhanced with the ant colony optimization method (original + ACO). The last one is based on the consolidation of word embeddings and original features (original + WE). The d/2 number of space extension parameter is adjusted to extend feature space due to its superior performance as stated in [3]. While the first half of features are original features, the remaining half is composed of significant features chosen with ACO, IG, and WE for the ACO-based, IG-based, and WE-based extended feature spaces, respectively. Our proposed approach is described in detail below.

WE-based features need some operations to consolidate with the original features while ACO-based and IG-based features are added to the end of feature space, directly. At first, d/2 number of features are randomly selected to obtain word embedding feature vector which includes the similarity measures of meaningfully related or surrounding words of actual word. After getting similarity vectors of d/2 number of randomly selected features, the best similarity score is chosen and divided into the total score of similarity vector to associate with the original feature space. This procedure mentioned above is repeated for all randomly selected features until we get d/2 number of new features to be added to the original feature space.

Algorithm 1. Extended Space Algorithm

Given: E={x_p, y_p}$_{p=1...N}$ =[X Y] where X is an N*d matrix including the training set and Y is an N dimensional column vector covering the class labels. d is the number of features, N is the number of training samples, T is the number of base learners, BL_i is the base learner, E_i is the extended training set for BL_i, EA is an ensemble algorithm.

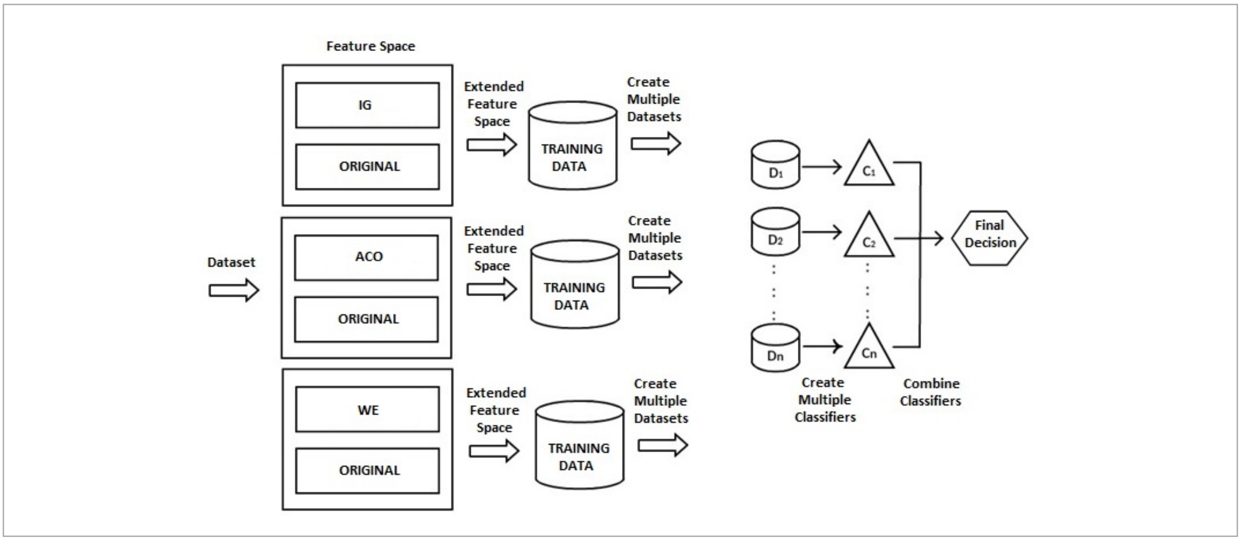
Initialization: Choose ensemble size T, the base learner model BL_i, and the ensemble algorithm EA.

Training:

```
for i=1:T
    1. Create new features ($EX_i$) by using feature selection techniques ($IG$, $ACO$), and word embeddings ($WE$).
    Generate d/2 number of features with $IG$ and store in $R_i$ or
    Generate d/2 number of  features with $ACO$ and store in $S_i$ or
    Generate d/2 significant features with $WE$ and store in $W_i$.
        Choose $d/2$ features, randomly.
        for w=1: d/2
    Create similarity vector and store in $SV_w$. Obtain the best similarity score from $SV_w$ and divide it by the total score of similarity vector. Then, store in $W_i$.
j=1
for z=1:d step by 2
    Create the jth new feature adding significant features with the proposed methods to $X$ matrix.
j=j+1
endfor
    2.Construct the new training set ($E_i$) by concatenating the matrix $X$ (original features) and $R_i$, or $X$ and $S_i$, or $X$ and $W_i$, seperately as $E_i$ =[$X$ $R_i$ $Y$], $E_i$ =[$X$ $S_i$ $Y$], $E_i$ =[$X$ $W_i$ $Y$], respectively.
    3. Train $BL_i$ with $E_i$ according to $EA$.
endfor
```

Figure 1

The process of extended feature space with our proposed technique



```
Testing:
for i=1:T
    1.Extend the feature space of the test sample.
    2.Classify the extended test sample with $BL_i$.
endfor
    Combine the base learners' decisions by the combination rule of the chosen ensemble algorithm $EA$.
```

After constructing the enriched feature space, conventional machine learning algorithms such as multinomial naïve Bayes, multivariate naïve Bayes, support vector machine, and random forest are performed to select baseline classifier for the proposed ensemble system. At the next step, ensemble strategy is carried out to maintain diversity and to obtain final decision of the system. Figure 1 illustrates the process of extended feature space with our proposed technique.

3.5. Ensemble of Classifiers

Ensemble algorithms used in this work are briefly mentioned. Bagging [3, 8, 18, 23, 27, 33, 36] generates new bootstrap samples utilizing substitution from the original dataset. Then, training is implemented on these samples. After that, the majority voting is utilized as an ensemble strategy. Random Subspace [3, 16, 18, 19, 25, 27, 13, 33, 37] exploits fairly simple randomness approach for the feature selection. Training is done with a subset of the original feature space instead of including all features for each base learner in the ensemble. Then, the classifier is constructed on different feature subsets illustrated randomly from the original feature set and associated by applying the majority voting. Random Forest [2, 3, 9] combines two approaches namely, Bagging and Random Subspace algorithms. Majority voting is employed for all ensembles to combine the decisions of base learners.

4. Experiment Setup

We have processed five different English datasets in our experiments. The first two datasets (Sts-Gold and Sts-Test) are utilized in the same way as described in [28]. Sts-Gold is manually labeled and a subset of tweets are chosen from the Standford Twitter Sentiment Corpus [15] and is presented by [28]. It contains 13 negative, 27 positive, and 18 neutral entitites as well as 1,402 negative, 632 positive, and 77 neutral tweets. It includes independent sentiment labels for tweets and entities, supporting the evaluation of tweet-based Twitter sentiment analysis models. The Standford Twitter Sentiment Corpus [15] consists of two different sets, training and test. Sts-Test is the test set of the Standford Twitter Sentiment Corpus. It is also manually annotated and encloses 177 negative, 182 positive, and 139 neutral tweets. Although the Sts-Test dataset is relatively small, it has been widely used in literature [5, 7, 15, 29, 30, 32] in different evaluation tasks.

The last three datasets are publicly available and gathered from Twitter in the second half of 2014. These are three real English, public and non-encoded datasets. Each dataset was labeled as positive or negative, according to the opinion expressed in respect to the object of interest. They are publicly available at http://www.dt.fee.unicamp.br/~tiago//sentcollection/. We evaluate our models by focusing on positive and negative tweets similar to the state-of-the-art studies [5, 15, 22, 29, 30, 32]. The class distribution of and main theme of datasets, when no preprocessing is applied, are summarized in Table 1. We don't apply any stemming or stop word filtering in order to avoid any bias that can be introduced by stemming algorithms or stop-word lists. Moreover, Sts-Gold dataset has an imbalanced class distribution. This is a well-known fact that machine learning algorithms are sensitive to an imbalanced class distribution. We also observe the impact of imbalance class distribution on the performance of proposed system. Experiments are carried out by modifying the training set levels and utilizing 5%, 10%, 30%, 50% and 80% percentages as the training data. The F-measure and accuracy percentage levels are abbreviated with "ts" affix to head a commotion off. The algorithms are launched at each training set levels by partitioning 10 parts randomly and stratified sampling is exploited at this step.

We have performed a statistical analysis evaluating Student's t-test to ensure that results were not obtained by chance. Significance level is set to 0.05 and the difference is accounted as statistically significant when the association of probability and Student's t-test is lower. The number of base learners is adjusted to 100 as represented in [1, 3]. As we mentioned before, feature extension parameter is set to d number of features for all datasets for comparing experiment results with impressive work [3]. To combine the decisions of base learners, majority voting is employed for all ensembles. By means of the most meaningful 100 features obtained by the information gain method, the feature space has been extended by varying the number of features in each data set. That is, the feature space of a dataset with a feature number of 50 is extended using 50 of the 100 most significant features obtained through information gain technique.

Table 1

Statistics of the datasets with no preprocessing

Dataset	#Positive	#Negative	Total	Theme
Sts-Gold	632	1402	2034	Misc.
Sts-Test	182	177	359	Misc.
Iphone6	371	161	532	Smartphone
Archeage	724	994	1718	Game
Hobbit	354	168	522	Movie

Moreover, it is necessary to specify some parameters for ACO feature selection process. First, the number of ants is equal to the number of features for each dataset. Because of this, the number of ants varies according to the dataset. Then, the algorithm has carried out a certain number of times. This is the same as the number of base learners, i.e. 100 times. After the algorithm has executed 100 times, the pheromone density is updated and a new set of ants are composed and the process iterates once more. The initial pheromone density of each feature is set to 1 at first. Two important information, local and global, about the traversal of ants are determined with the parameters α and β. The choice of α and β is specified experimentally and set to 1 and 0.1, respectively. The pheromone trail evaporation coefficient (ρ=0.2) is a parameter to update pheromone trails and located in the range between 0 and 1.

We utilize open source machine learning software which is called WEKA for the feature selection process. The proposed extended feature space system is constructed on this software with Java programming language. Besides, this work employs the Python 3 version of word2vec in the Gensim theme model with Pycharm environment, which only carries out the continuous skip-gram model and trains with the hierarchical softmax method.This model utilizes a 200-dimensional vector space to demonstrate words and the training window is set to 5. Moreover, Google has used Google News dataset that contains about 100 billion words to obtain pre-trained vectors with the Word2Vec Skip-gram algorithm [12, 24]. The pre-trained model includes word vectors for about 3 million words and phrases. We use this pre-trained model in English to represent documents with 200 dimensions or features.

5. Experiment Results

The conducted experiments demonstrate the short sentiment classification success of each baseline classifier over five datasets in Table 2. Bold values demonstrate the best scores. F-measure and accuracy results are utilized as evaluation metric to demonstrate the contribution of our work. Abbreviations are employed as follows: BG: Bagging, BS: Boosting, RS: Random subspaces, RF: Random forest, X_{IG}: Extended feature space with IG-based features for X ensemble algorithm, X_{ACO}: Extended feature space with ACO-based features for X ensemble algorithm and X_{WE}: Extended feature space with WE-based features for X ensemble algorithm.

Table 2

Averaged F-measure results of each baseline classifier at ts80

Dataset	MNB	MVNB	SVM	RF
Sts-Gold	82.15±0.07	81.36±0.04	**83.44±0.02**	82.90±0.06
Sts-Test	81.30±0.05	80.12±0.02	**82.96±0.01**	81.75±0.04
Iphone6	70.42±0.03	**74.48±0.05**	73.66±0.03	72.15±0.09
Archeage	85.13±0.02	85.91±0.05	**86.20±0.03**	84.30±0.04
Hobbit	87.10±0.04	84.36±0.02	**90.45±0.02**	88.23±0.08
avg	81.22±0.04	81.25±0.03	**83.34±0.02**	81.87±0.06

As it can be seen in Table 2, the best F-score performance is achieved by SVM by assessing averaged F-score values of each baseline classifier. RF has a slightly better performance than MNB and MVNB while MNB and MVNB have almost the same classification success. Hence, SVM as a base learner will be a good choice in terms of classification performance because of the highest F-measure values. Eventually, the classification success of the base learners is ordered as SVM > RF > MVNB > MNB.

Table 3

Averaged F-measure results of the combination of ensemble algorithms and SVM baseline classifier on original data at ts80

Dataset	SVM	BG_o	BS_o	RS_o	RF_o
Sts-Gold	83.44	83.47	**83.70**	83.65	82.90
Sts-Test	82.96	82.51	82.94	**83.03**	81.75
Iphone6	73.66	73.82	74.05	**74.18**	72.15
Archeage	86.20	85.26	85.48	**86.43**	84.30
Hobbit	**90.45**	89.82	90.17	90.06	88.23
avg	83.34	82.98	83.27	**83.47**	81.87

In Table 3, the F-measure results of ensemble algorithms are represented on the original data when the baseline classifier is set to SVM. By observing the system performance of the only ensemble algorithms on the original data, we can make sure that the extended feature spaces based classifier ensembles are worth improving system success. It is clearly seen that the baseline classifier SVM generally performs well without using any ensemble algorithm. The combination of baseline classifier and random subspace as an ensemble method exhibit better success than the baseline classifier in view of the fact that the averaged F-measure results are considered. The success of homogeneous classifier ensembles on original data is summarized as RS_o > SVM > BS_o > BG_o > RF_o even if averaged F-measure results are very close each other except RF.

The results demonstrate that the proposed WE-based ensemble systems evidently present an overall superior performance to any of the other evaluated extended feature space based ensemble system in Table 4. The classification success is ordered as RS_{WE} > BS_{WE} > BG_{WE} > RS_{ACO} > BS_{ACO} > BG_{ACO} > RS_{IG} > BS_{IG} > BG_{IG} > SVM at ts80. All versions of the enhanced space based ensemble systems significantly contribute to the classification performance by improving up to 5% compared to the baseline classifiers. The performance of ensemble algorithms is RS > BS > BG for all extended space versions in terms of averaged F-measure results. Moreover, space is extended with word embedding based features due to its superior success. The classification success of extension techniques is ordered as WE > ACO > IG for all datasets when the ensemble algorithm is set to RS. The classification performance of IG and ACO-based extended feature space is competitive but not enough to claim statistically significant because of the closeness of results in terms of ensemble algorithms. Thus, the combination of random subspace as an ensemble algorithm and WE-based extended feature space yields by far the highest results at ts80. In other words, our proposed method with 88.76% result (RS_{WE}) is the best model to enhance the classification performance for all datasets in terms of averaged F-measure results.

When the averaged F-measure results of Table 3 and Table 4 are compared, extended spaces based classi-

Table 4

Averaged F-measure results of the proposed method on extended feature spaces at ts80

Method	SVM	BG_{IG}	BS_{IG}	RS_{IG}	BG_{ACO}	BS_{ACO}	RS_{ACO}	BG_{WE}	BS_{WE}	RS_{WE}
Sts-Gold	83.44	83.40	83.45	83.88	83.72	83.91	84.20	86.46	86.95	**88.44**
Sts-Test	82.96	82.80	82.91	83.14	82.95	83.12	83.77	85.90	86.53	**87.45**
Iphone6	73.66	74.10	74.25	74.40	74.75	74.82	75.10	77.23	78.67	**79.95**
Archeage	86.20	86.53	86.70	86.80	86.75	86.90	87.05	89.21	90.23	**91.55**
Hobbit	90.45	90.12	90.20	90.55	90.44	90.73	91.33	94.22	95.88	**96.41**
avg	83.34	83.39	83.50	83.75	83.72	83.90	84.29	86.60	87.65	**88.76**

fier ensembles always exhibit superior performance compared to the combination of baseline classifier and ensemble algorithms on the original data in terms of ensemble methods. The combination of the IG-based and ACO-based extended spaces and bagging as an ensemble algorithm performs almost 1% better classification performance than the BG_o which is the combination of baseline classifier and bagging algorithm on the original data. Moreover, the performance enhancement of the consolidation of WE-based extended space and bagging method (86.60%) nearly reaches 4% in comparison to the BG_o with 82.98% classification success. To summarize, the classification success is ordered as BG_{WE} (86.60%) > BG_{ACO} (83.72) > BG_{IG} (83.39) > BG_o (82.98) in terms of bagging algorithm. This order of success also applies for the other ensemble algorithms when extended space and original versions of data are considered. These noticeable results clearly demonstrate the contribution made by the combination of the ensemble algorithms with extended feature spaces to overall system performance.

In Figure 2, the classification performance of extension techniques is presented by varying the number of base learners when the ensemble algorithm is adjusted to RS. It is obviously seen that the number of base learners which varies from 10 to 150 is also a significant measure to observe the classification success of extended space based techniques. When the number of base learners is raised up to 100, the accuracy results also boost for each extended space method. However, as the number of base learners continues to increase after 100, the classification success considerably decreases as inversely proportional. For this reason, it is determined as 100 in experiments and these results are consistent with the literature [3, 20].

In Figure 3, classification performance of random subspace algorithm is evaluated in terms of the proposed extended spaces. WE-based extended space model outperforms traditional feature selection techniques at all training set percentages. The difference of results between WE-based and IG-based models is prominently observed up to 10% at smaller training set sizes. On the other hand, ACO-based extended space model is more competitive than IG-based model in terms of averaged accuracy results. The ACO-based model achieves an improvement of 2% more than the IG based model. Thus, the difference between the ACO-based model and the WE-based model reaches a maximum of 8% at smaller training set percentages. Although the ACO-based model exhibits about 1% enhancement compared to the IG-based model at ts80, the classification success of two models is generally very close to each other at all training set levels. As a result, the best technique to extend the feature space is to integrate word embedding based features (WE) and original ones.

Figure 2

Accuracy results of feature space extension techniques

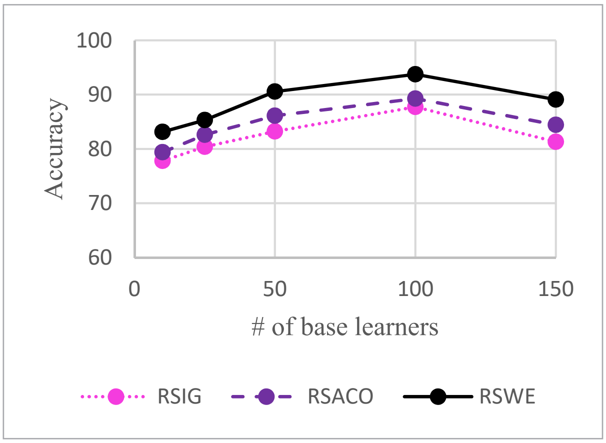
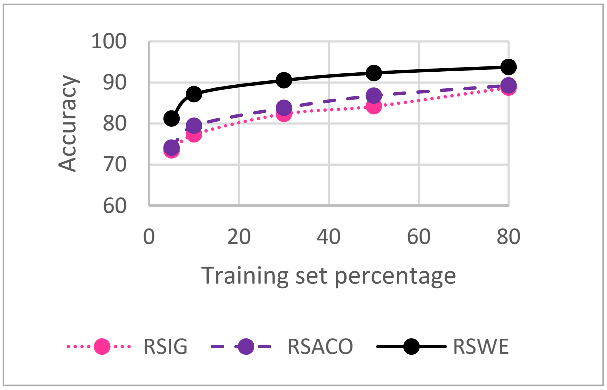


Figure 3

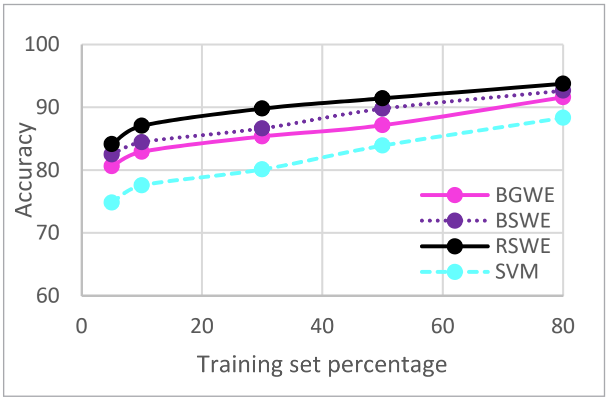
Accuracy results of RS ensemble algorithm in terms of the proposed extended spaces



In Figure 4, the classification success of SVM as a base learner and WE-based extended spaces are analyzed in terms of ensemble algorithms. Extending the feature space with the WE-based model and using ensemble algorithms ensures minimum 3% - maximum 10% improvement compared to the classification per-

Figure 4

Accuracy results of WE-based extended spaces in terms of ensemble algorithms



formance of base learner at ts80 and ts5, respectively. That is, the difference between the proposed models and base learner increases at smaller percentages and vice versa. As the training set percentage decreases, the difference between the classification success of each ensemble algorithm also becomes more pronounced. For example, the BS method demonstrates 1% better performance than BG while the RS model present 1% better enhancement compared to BS at ts80. RS and BS improve the classification performance of system by approximately 4% and 2%, respectively, compared to the BG at smaller percentages.

As a result, extending feature space with word embedding based techniques and combining this feature space with ensemble algorithms presents much better classification performance compared to a single classifier. Figures 3 and 4 clearly show that the feature space extended by word embedding based features in short text classification will provide the best classification performance by diversifying random subspace as an ensemble algorithm.

It is considerable to compare experiment results with the state-of-the-art studies [4, 22] on extended spaces to demonstrate the contribution of our proposed technique. Lochter et al. [22] employ nine datasets and four of them are common with ours as seen in Table 5. The superiority of our proposed method is obviously observed for all datasets when the same training settings are adjusted.

Table 5

Comparison of the classification success of our proposed method vs study1 [22] in terms of F-measure results

Method	Sts-Test	Iphone6	Archeage	Hobbit
RS_{WE}	***87.4***	***79.9***	***91.5***	***96.4***
Study1	86.3	73.8	86.9	92.1

The other study [4] proposes nine different methods for the ensemble system. Araque et al. [4] employ six datasets and one of them is common with ours as seen in Table 6. Our proposed method outperforms all methods of them except MSG+bg technique. The slightest difference between ours (88.4%) and theirs (89.2%) can be arisen from differences in experimental settings and not be enough to claim statistically considerable because of the closeness of averaged F-scores. Thus, it is noteworthy to specify that the combination of our proposed technique (RSWE) predominantly surpasses state-of-the-art studies.

Table 6

Comparison of the classification success of our proposed method vs study2 [4] in terms of F-measure results

Method	Sts-Gold
RS_{WE}	***88.4***
M_{G}	83.4
CEM_{SG}^{Vo}	83.5
CEM_{SG}^{ME}	84.5
M_{SG}	84.7
M_{SG+bg}	***89.2***
M_{GA}	85.2
M_{SGA+bg}	85.2
CEM_{SGA}^{Vo}	87.0
CEM_{SGA}^{ME}	85.5

6. Discussion and Conclusion

The superiority of ensemble systems is a widely accepted assumption in machine learning domain as mentioned before. Owing to this approach, it is recommended to produce more accurate and robust models. In this work, we propose to investigate the contribution of extended spaces to the classification performance by employing ensemble algorithms. For this purpose, we take the concept one step further in extended spaces by utilizing word embedding based features (WE) which have not been consolidated before. Moreover, this is the first research for the extended spaces with classifier ensembles in terms of using both traditional feature selection techniques and word embedding based feature extraction method. Features chosen with IG, ACO, and WE are blended with the original features to constitute a new extended feature space. Then, the enriched feature space is carried out on three popular ensemble algorithms (bagging, boosting, and random subspaces) by utilizing SVM as a baseline classifier. Finally, the extended spaces developed by our proposed method maintain noteworthy enhancement to the classification performance in comparison to the original version and various extended versions of recent state-of-the-art studies. Considering the overall classification performances, feature spaces with the original ones have the lowest classification performance at all training set levels and this is an indicator that the original feature spaces need to develop.

As well as the classification success of proposed system, the analysis of execution time is also evaluated in terms of training time. More training time is needed for the enhanced spaces compared to the original ones. Because, the enhanced space features effect the search time of the features owing to covering more features. The training time for Twitter corpus is around 1h25min using 12 threads in an Intel® Xeon® E5-2643 3.30 GHz machine. We consider that performing our proposed model for a GPU environment can have a great influence on the training time performance.

To sum up, the extended spaces with our proposed approach advance the classification success of system compared to the original versions. Over and above, it is observed that the enhancement of extended spaces with classifier ensembles using word embeddings exhibits better classification performance in comparison to the other enhanced space techniques. In future, we plan to apply different base learners to the classification problems.

References

1. Adnan, M. N., Islam, M. Z. Comprehensive Method for Attribute Space Extension for Random Forest. International Conference on Computer and Information Technology, Dhaka, Bangladesh, December 22-23, 2014, 25-29. https://doi.org/10.1109/ICCITechn.2014.7073129
2. Adnan, M. N., Islam, M. Z., Kwan, P. W. H. Extended Space Decision Tree. International Conference on Machine Learning and Cybernetics, Lanzhou, China, July 13-16, 2014, 219-230. https://doi.org/10.1007/978-3-662-45652-1_23
3. Amasyalı, M. F., Ersoy, O. K. Classifier Ensembles with the Extended Space Forest. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(3), 549-562. https://doi.org/10.1109/TKDE.2013.9
4. Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., Iglesias, C. A. Enhancing Deep Learning Sentiment Analysis with Ensemble Techniques in Social Applications. Expert Systems and Applications, 2017, 77, 236-246. https://doi.org/10.1016/j.eswa.2017.02.002
5. Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., Varma, V. Mining Sentiments from Tweets. Proceedings of 12th International Conference on Wireless Algorithms, Systems, and Applications, Guilin, China, June 19-21, 2017, 11-18.
6. Becker, W., Wehrmann, J., Cagnini, H. E. L., Barros, R. C. An Efficient Deep Neural Architecture for Multilingual Sentiment Analysis in Twitter. Proceedings of 30th International Conference on Florida Artificial Intelligence Research Society, Marco Island, Florida, May 22-24, 2017, 246-251.
7. Bravo-Marquez, F., Mendoza, M., Poblete, B. Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis. Proceedings of the 2nd International Workshop on Issues of Sentiment Discovery and Opinion Mining, Chicago, IL, USA, August 11-14, 2013, 1-9. https://doi.org/10.1145/2502069.2502071
8. Breiman, L. Bagging Predictors. Machine Learning, 1996, 24, 123-140. https://doi.org/10.1023/A:1018054314350
9. Breiman, L. Random Forests. Machine Learning, 2001, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
10. Chaurasia, V., Pal, S. Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. Interna-

tional Journal of Computer Science and Mobile Computing, 2014, 3(1), 10-22.

11. Chen, Q., Guo, Z., Sun, C., Li, W. Research on Chinese Micro-Blog Sentiment Classification Based on Recurrent Neural Network. Proceedings of 2nd International Conference on Computer Science and Technology, Guilin, China, June 26-28, 2017, 859-867. https://doi.org/10.12783/dtcse/cst2017/12594

12. English Pre-trained Word2vec Model, https://code.google.com/archive/p/word2vec/. Accessed on February 10, 2018.

13. Garcia-Pedrajas, N., Ortiz-Boyer, D. Boosting Random Subspace Method. Neural Networks, 2008, 21(9), 1344-1362. https://doi.org/10.1016/j.neunet.2007.12.046

14. Ghosal, D., Bhatnagar, S., Akhtar, M. S., Ekbal, A., Bhattacharyya, P. IITP at Semeval-2017 task 5: An Ensemble of Deep Learning and Feature Based Models for Financial Sentiment Analysis. Proceedings of 11th International Workshop on Semantic Evaluations, Vancouver, Canada, August 3-4, 2017, 899-903. https://doi.org/10.18653/v1/S17-2154

15. Go, A., Bhayani, R., Huang, L. Twitter Sentiment Classification Using Distant Supervision. CS224N Project Report, Stanford, 2009, 1(12), 1-6.

16. Ho, T. K. The Random Subspace Method for Constructing Decision Forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8), 832-844. https://doi.org/10.1109/34.709601

17. Hu, F., Li, L., Zhang, Z., Wang, J., Xu, X. Emphasizing Essential Words for Sentiment Classification Based on Recurrent Neural Networks. Journal of Computer Science and Technology, 2017, 32(4), 785-795. https://doi.org/10.1007/s11390-017-1759-2

18. Kotsiantis, S. Combining Bagging, Boosting, Rotation Forest and Random Subspace Methods. Artificial Intelligence Review, 2011, 35, 223-240. https://doi.org/10.1007/s10462-010-9192-8

19. Kuncheva, L. I., Rodríguez, J. J., Plumpton, C. O., Linden, D. E. J., Johnston, S. J. Random Subspace Ensembles for fMRI Classification. IEEE Transactions on Medical Imaging, 2010, 29(2), 531-542. https://doi.org/10.1109/TMI.2009.2037756

20. Kuncheva, L. I., Whitaker, C. J. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. Machine Learning, 2003, 51(2), 181-207. https://doi.org/10.1023/A:1022859003006

21. Liao, S., Wang, J., Yu, R., Sato, K., Cheng, Z. CNN for Situations Understanding Based on Sentiment Analysis of Twitter Data. Procedia Computer Science, 2017, 111, 376-381. https://doi.org/10.1016/j.procs.2017.06.037

22. Lochter, J. V., Zanetti, R. F., Reller, D., Almeida, T. A. Short Text Opinion Detection Using Ensemble of Classifiers and Semantic Indexing. Expert Systems and Applications, 2016, 62, 243-249. https://doi.org/10.1016/j.eswa.2016.06.025

23. Martin, B., Marsík, J. Multi-Label Text Classification via Ensemble Techniques. International Journal of Computer and Communication Engineering, 2012, 1(1), 62-71. https://doi.org/10.7763/IJCCE.2012.V1.18

24. Mikolov, T., Chen, K., Corrado, G. Dean, J. Efficient Estimation of Word Representations in Vector Space. September 7, 2013, 1-12.

25. Nanni, L., Lumini, A. Random Subspace for an Improved Biohashing for Face Authentication. Pattern Recognition Letters, 2008, 29(3), 295-300. https://doi.org/10.1016/j.patrec.2007.10.005

26. Nozza, D., Fersini, E., Messina, E. Deep Learning and Ensemble Methods for Domain Adaptation. Proceedings of 28th International Conference on Tools with Artificial Intelligence, California, USA, November 7-9, 2011, 184-189. https://doi.org/10.1109/ICTAI.2016.0037

27. Panov, P., Dzeroski, S. Combining Bagging and Random Subspaces to Create Better Ensembles. International Conference on Intelligent Data Analysis, Berlin, Germany, September 6-8, 2007, 118-129. https://doi.org/10.1007/978-3-540-74825-0_11

28. Saif, H., Fernandez, M., He, Y., Alani, H. Evaluation Datasets for Twitter Sentiment Analysis: A Survey and a New Dataset, the sts-Gold. International Workshop on Emotion and Sentiment in Social and Expressive Media, Turin, Italy, December 3rd, 2013, 2-14.

29. Saif, H., He, Y., Alani, H. Semantic Smoothing for Twitter Sentiment Analysis. Proceedings of the 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, 16-20.

30. Saif, H., He, Y., Alani, H. Alleviating Data Sparsity for Twitter Sentiment Analysis. Proceedings, 2nd Workshop on Making Sense of Microposts, Lyon, France, April 16, 2012, 2-9.

31. Santos, C. N., Gatti, M. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. Proceedings of 25th International Conference on Computational Linguistics, Dublin, Ireland, August 23-29, 2014, 69-78.

32. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. Proceedings of the 1st

Workshop on Unsupervised Learning in Natural Language Processing, Edinburgh, Scotland, July 30, 2011, 53-63.

33. Skurichina, M., Duin, R. P. W. Bagging, Boosting and the Random Subspace Method for Linear Classifiers. Pattern Analysis and Applications, 2002, 5, 121-135. https://doi.org/10.1007/s100440200011

34. Uysal, A. K., Gunal, S. The Impact of Preprocessing on Text Classification. Information Processing and Management, 2014, 50(1), 104-112. https://doi.org/10.1016/j.ipm.2013.08.006

35. Uysal, A. K., Murphey, Y. L. Sentiment Classification: Feature Selection Based Approaches Versus Deep Learning. IEEE International Conference on Computer and Information Technology, Helsinki, Finland, August 21-23, 2017, 23-30. http://doi.ieeecomputersociety.org/10.1109/CIT.2017.53

36. Wang, X., Tang, X. Random Sampling for Subspace Face Recognition. International Journal of Computer Vision, 2006, 70, 91-104. https://doi.org/10.1007/s11263-006-8098-z

37. Wang, G., Zhang, Z., Sun, J., Yang, S., Larson, C. A. POS-RS: A Random Subspace Method for Sentiment Classification Based on Part-Of-Speech Analysis. Information Processing and Management, 2015, 51(4), 458-479. https://doi.org/10.1016/j.ipm.2014.09.004

38. Young, T., Hazarika, D., Poria, S., Cambria, E. Recent Trends in Deep Learning Based Natural Language Processing. IEEE Computational Intelligence Magazine, 2018, 13(3), 55-75. https://doi.org/10.1109/MCI.2018.2840738

39. Zhao, Z., Lu, H., Cai, D., He, X., Zhuang, Y. Microblog Sentiment Classification via Recurrent Random Walk Network Learning. Proceedings of 26th International Conference on Artificial Intelligence, Melbourne, Australia, August 19-25, 2017, 3532-3538. https://doi.org/10.24963/ijcai.2017/494

40. Zheng, Z., Wu, X., Srihari, R. Feature Selection for text Categorization on Imbalanced Data. SIGKDD Explorations, 2004, 6(1), 80-89. https://doi.org/10.1145/1007730.1007741