IWD Based Feature Selection Algorithm for Sentiment Analysis

Tuba Parlar[1], Esra Sarac[2]

[1]*Department of Computer Technologies, Hatay Mustafa Kemal University,*
Hatay, Turkey
[2]*Department of Computer Engineering, Adana Science and Technology University,*
Adana, Turkey
tparlar@mku.edu.tr

Abstract—**Feature selection methods aim to improve the classification performance by eliminating non-valuable features. In this paper, our aim is to apply a recent optimization technique namely the Intelligent Water Drops (IWD) algorithm to select best features for sentiment analysis. We investigate the classification performances of our proposed IWD based feature selection method by comparing one of the well-known feature selection method using Maximum Entropy classifier. Experimental results show that Intelligent Water Drops based feature selection method outperforms than ReliefF method for sentiment analysis.**

Index Terms—**Feature selection; Machine learning; Natural language processing; Text mining; Sentiment analysis.**

I. INTRODUCTION

Today internet and social media have become an important source of information. The increasing amount of information on the internet has brought new research areas. Opinions are shared by a lot of people through different social media platforms such as blogs, forums, twitter. With the increasing importance of sentiment information, there is a need for fast and effective analysis techniques that are not only subject-oriented but also sentiment focused [1]. However, the rapidly increasing amount of online documentation has made it difficult and time-consuming to analyze desired information. Sentiment analysis has become an important research area for automatic analysis of review documents. Sentiment analysis aims to classify the sentiments of these review documents especially into two classes: positive or negative. In order to increase the performance of the classification process, feature selection methods are applied to determine the most valuable features. Feature selection is very important in two respects: the efficiency of the training process increases significantly by reducing the number of features; secondly, the accuracy of classification increases by choosing most valuable features.

In this study, our aim is to investigate the effects of our proposed IWD based feature selection algorithm on classification performance for sentiment analysis. For this purpose, we compared our IWD based feature selection method with a well-known feature selection method called

ReliefF. This paper is organized as follows: in the next section, we introduce the relevant studies in the literature. In the third section, we describe the methods that we use for selecting features, present the information about the data, performance evaluation for our experiments. In the fourth section, we present the experimental results and finally we present the main conclusion.

II. RELATED WORK

Sentiment analysis can be defined as a text classification process based on the classification of review documents as generally positive or negative. Machine learning methods [2] or lexicon based methods [3] have been widely used for sentiment classification. However, machine learning methods have been preferred by many researchers because of their ease of use and computational performance [2]–[4]. Naive Bayes (NB), Support Vector Machines (SVM) and Maximum Entropy (ME) classifiers have been applied with success to sentiment classification problems.

Feature selection methods rank features so that non-informative features can be eliminated in order to improve the classification accuracy and efficiency. Researchers have used various feature selection methods to distinguish features that contain more information [5]–[7]. Agarwal and Mittal [7] employ information gain (IG) and minimum redundancy maximum relevancy (mRMR) feature selection methods. They show that mRMR method obtains better classification performance using boolean multinomial NB classifier. Abbasi, Chen, and Salem [5] compare information gain method with an intuitive method namely Entropy Weighted Genetic Algorithm (EWGA) and EWGA achieves better accuracy using SVM classifier. Stylios, Katsis, and Christodoulakis [8] compare two heuristic method namely Ant Colony Optimization (ACO) and Particle Swarm Optimization for feature selection and they improve the accuracy from 83.7 % to 90.6 %. Kaur, Sehra, and Sehra [9] investigate the classification performance of a hybrid method that combines SVM and ACO, and they achieve to improve the accuracy from 75.5 % to 86.6 %. Some researchers propose a modified IWD algorithm with alternative selection methods namely linear ranking and exponential ranking [10]. Moncayo-Martinez and Mastrocinque [11] propose an IWD based optimization method to design and configure a supply chain and logistics

network taking into account multiple objectives simultaneously. In another study [12], researchers apply an algorithm with IWD to select nodes for service-oriented wireless networks. Moreover, Alexandros and Georgios [13] underline the importance of nature inspired optimization algorithms and present recent developments and possibilities of using for different problem areas of these algorithms in terms of theory and application.

III. METHODS AND DATASET

In this study, we implemented an IWD based feature selection method and compared its effects on classification performance for sentiment analysis with ReliefF feature selection method. While ReliefF is a statistical based method, IWD based method is a heuristic method.

A. ReliefF

Relief algorithm is proposed by Kira and Rendell [14] as a simple, fast, and effective approach to feature weighting. Kononenko [15] extends Relief algorithm and proposes the ReliefF algorithm for multi class problems. ReliefF estimates $W[A]$ of the quality of attribute A according the equation in line 8–9 in Fig. 1 [16]. In Fig. 1, n indicates the number of training instances, a indicates the number of features and m indicates the number of random training instances out of n used to update W. We use RelifF algorithm from Weka data mining software package [17].

Algorithm ReliefF
Input: for each training instance, a vector of feature values and the class value
1. initialize vector W
2. **for** i= 1 to m **do**
3.   randomly select a target instance R_i;
4.   find a nearest hits H and nearest miss M.
5.   **for** A= 1 to a **do**
6. $W[A] = W[A] - diff\ (A, R_i, H)/m + diff\ (A, R_i, M)/m$
7.   **end**
8. **end**
9. **return** W

Fig. 1. Pseudo code of ReliefF feature selection algorithm.

When performing weight updates, the difference in value of attribute A between two instances I_1 and I_2, where $I_1 = R_i$ and I_2 is either H or M is calculated by *diff* function in Fig. 1. [16]. *diff* function is defined as follows for discrete features

$$diff\left(A, I_1, I_2\right) = \begin{cases} 0, & if\ value\left(A, I_1\right) - value\left(A, I_2\right), \\ 1, & otherwise. \end{cases} \quad (1)$$

For continuous features, *diff* function is defined as follows

$$diff\left(A, I_1, I_2\right) = \frac{value\left(A, I_1\right) - value\left(A, I_2\right)}{\max\left(A\right) - \min\left(A\right)}. \quad (2)$$

The *max(A)* and *min(A)* values determined over the whole set of instances. By this normalization, all weight updates fall between 0 and 1 for all type of features. When updating $W[A]$, to normalize final weights between -1 and 1, output of *diff* function is divided by m.

B. Intelligent Water Drops

IWD based feature selection algorithm [7] is implemented to select best features to provide an accurate and fast classification. The IWD algorithm constructs an optimal solution through cooperation among a group of agents called water drops. The algorithm imitates the phenomena of a swarm of water drops flowing with soil along a river bed. Procedurally, each water drop incrementally constructs a solution through a series of iterative transitions from one node to the next until a complete solution is obtained. Water drops communicate with each other through an attribute called soil, which is associated with the path between any two points. The soil value is used to determine the direction of movement from the current node to the next, whereby a path with a lower amount of soil is likely to be followed. The algorithm of IWD based feature selection are described in pseudo code given below:

Algorithm IWD based feature selection
Input: set of all features
Output: set of best n features
1. Initialize parameters
2. repeat
3. **for each** IWD **do**
4.   choose n features by using roulette wheel that is formed according to the probabilities of features that are computed with respect to soil values of features.
5.   classify documents by using selected n features and compute the F-score.
6.   by using the computed F-score value, update the soil and velocity values of the IWD, and the soil value of all features in the feature space.
7. **end**
8. choose the best feature set found so far.
9. **until** the termination condition is met.
10. **return** the best feature set.

Fig. 2. Pseudo code of IWD based feature selection algorithm.

First, we construct a graph from our features to use IWD algorithms and then we set the initial soil (*initSoil*) of features, initial velocity of water drops (*initVel*), number of water drops (N_{IWD}), soil updating parameters (a_s, b_s, c_s, ρ), and velocity updating parameters (a_v, b_v, c_v). After initialization the parameters, for each IWD, we choose n (distinct) features (f). We compute the probability $P(i)$ according to (3). At the beginning, water drops are spread randomly at the nodes of the construction graph, and visited nodes list (F) is updated to include the start node

$$P(i) = \frac{f\left(soil(i)\right)}{\sum_{k \notin F} f\left(soil(k)\right)}, \quad (3)$$

where $P(i)$ represents the probability of selecting node i, $f(soil(i))$ is the fitness function. The fitness function of candidate node i is inversely proportional to the absolute soil value (4)

$$f\left(soil(i)\right) = \frac{1}{\varepsilon_s + g\left(soil(i)\right)}, \quad (4)$$

where ε_s is a small positive number used to prevent the division by zero in function $f\left(soil(i)\right)$

$$g\left(soil(i)\right) =$$

$$= \begin{cases} soil(i), & if \ \min_{k \notin F}\left(soil(k)\right) \geq 0, \\ soil(i) - \min_{k \notin F}\left(soil(k)\right), & otherwise, \end{cases} \quad (5)$$

where $soil(i)$ refers to the amount of soil within the node (feature) i and $soil(k)$ refers to the amount of soil within the node k which is not selected before. Then, add the newly selected node i to the list F. The min() function returns the minimum value of its arguments. After selecting n features according to (3), the training dataset is classified by using the selected features and the F-score value which is a real number in range [0, 1], of the classification is taken to measure the "quality" of the selected features. If the computed F-score is high, this means that the selected features are valuable. Then, F-score value is used to update the velocity and the soil values. For each IWD, the velocity is updated as follows

$$vel^{IWD}(t+1) = \frac{vel^{IWD}(t) + a_v}{b_v + c_v \times \sum_{i \in F} soil(i)}, \quad (6)$$

where a_v, b_v, c_v are the static parameters used to represent the nonlinear relationship between the velocity of water drop IWD, i.e. vel^{IWD}, and the inverse of the amount of soil in the local node, i.e. $soil(i)$. The amount of soil that the IWD loads from the selected features is

$$\Delta soil(F) = \frac{a_s}{b_s + c_s \times time^2\left(F, vel^{IWD}(t+1)\right)}, \quad (7)$$

where

$$time^2\left(F, vel^{IWD}(t+1)\right) = \frac{1}{Fscore(F) \times vel^{IWD}(t+1)}, \quad (8)$$

here F refers to the selected subset of features and $Fscore(F)$ refers to the F-score value of the selected features. The IWD's soil, $soil^{IWD}$, is increased by removing some soil of the selected features F. Update the soil of the IWD as

$$soil^{IWD} = soil^{IWD} + \Delta soil(F). \quad (9)$$

The soil values of all features are updated as follows

$$soil(i) =$$

$$= \begin{cases} (1-p) \times soil(i) - 2\rho \times \Delta soil(F), & if \ i \in F \ and \ F \ is \ best, \\ (1-p) \times soil(i) - \rho \times \Delta soil(F), & if \ i \in F \ and \ F \ is \ not \ best, \\ soil(i), & if \ i \notin F. \end{cases} \quad (10)$$

The above computations are repeated for each IWD and the best feature set is recorded. All these processes continue until the termination condition is met. We chose 250 iterations as the termination condition. According to the previous studies [18], [19] no improvement observed after 250 iterations for swarm based algorithms.

C. Sentiment Classification

Many researchers widely use machine learning methods because they can be easily adapted. Sentiment classification is a text classification task to sort the documents into negative and positive classes [20]. Maximum Entropy is also known as logistic regression in statistics [21]. Maximum Entropy is more robust to correlated features. If there are many correlated features, ME assigns more accurate probability. We use Logistic Regression classifier from Weka data mining software package [17].

D. Dataset

We conduct experiments on the Turkish Twitter dataset [22] belonging to a private telecommunication company. The dataset contains 3000 tweets that consists of three different classes (positive, negative, neutral). We use only positive and negative tweets. We tokenize alphabetic characters as features and calculate their weights using term frequency. Five-fold cross validation is applied to evaluate the performance of classification.

E. Evaluation Metrics

The classification performance of experiments is evaluated using F-score. F-score is based on precision and recall measurement terms. Precision (P) is the percentage of correctly classified documents among all documents that classified to a class. Recall (R) is the percentage of documents that classified to that class. F-score is defined as the harmonic mean of recall and precision [23]

$$F - score = 2 \times \frac{P \times R}{P + R}. \quad (11)$$

IV. EXPERIMENTS AND RESULTS

In this study, Turkish Twitter data is used and we investigate the effects of our proposed IWD based feature selection method by comparing traditional feature selection method for sentiment analysis. In a previous study on cyberbullying which is a special category of sentiment analysis, using emoticons as features is not effected classification performance well [24]. Therefore, we extract only alphabetic characters as features and calculate their weights using term frequency. We do not apply stemming because of Turkish is an agglutinative language and the suffixes include the polarity of a word [25]. We developed a software using Python Natural Language Toolkit-NLTK [26] for our experiments. We run Maximum Entropy classifier using five-fold cross validation. In the first stage of our study, we establish the baseline for examining the impacts of feature selection methods using all features that are obtained bag-of-words method. We use RelifF algorithm from Weka data mining software package. In the second stage, the features are ranked by each feature selection method. Our aim is to select the most valuable features. We chose top ranked 10, 50, 100, 250, and 500 number of features for measuring classification performances.

IWD parameters are initialized according to [27] as $initSoil = 1000$; $initVel = 100$; $a_v = 1000$; $b_v = 0.01$; $c_v = 1$; $a_s = 1000$; $b_s = 0.01$; $c_s = 1$; $\rho = 0.01$. Number of iterations and number of IWDs are determined according to previous studies [18], [19] as $N_{IWD} = 30$ and number of iterations = 250. Our IWD based feature selection algorithm selects a predefined number of features for each water drop. Then we

evaluate the performance of the selection made by each water drop by using the ME classifier. According to the performance of the selected features, we update soil and velocity values and selection probabilities of the features. This process is repeated until the best features are selected. After selecting the best features, classification of the test dataset is made by using the selected feature set.

We obtain our baseline as 0.691 by using all features in the training set (without feature selection) which can be observed in Table I. There are many non-informative features in the feature set and these non-informative features negatively affect the classification performance. Selecting feature subsets eliminates non-informative features and improves the classification performance. When we compare the results presented in Table I, we can say that classification performances increase in terms of F-score when we use features selected by our proposed IWD based feature selection method. For example, the F-score for the Twitter dataset is increased from 0.691 to 0.721 using IWD based feature selection method. Our proposed method achieved best performances with all feature sizes than ReliefF filter based method using Maximum Entropy classifier.

TABLE I. RESULTS OF THE FEATURE SELECTION METHODS WITH REDUCED FEATURE SIZES.

Feature Sizes	ReliefF Method	IWD Based Method
10	0.617	**0.643**
50	0.667	**0.697**
100	0.682	**0.701**
250	0.683	**0.721**
500	0.679	**0.707**
All features	**0.691**	

In Table II, time required to classify test dataset for different feature sizes with ME classifier are displayed. As can be easily seen from the Fig. 3, using feature selection methods reduces the time required to classify new (unseen) data sharply without making reduction in classification accuracy.
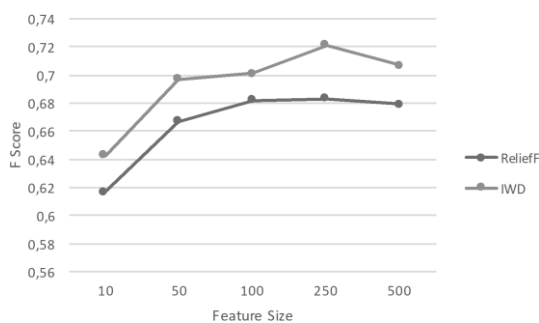

Fig. 3. Results of the feature selection methods with reduced feature sizes.

TABLE II. EFFECTS OF FEATURE SELECTION ON TIME NEEDED FOR CLASSIFICATION.

Feature Sizes	Classification time (sls)
10	24
50	36
100	61
250	168
500	189
All features	4554

V. CONCLUSIONS

In this study, we have developed an IWD based feature selection system for the sentiment classification. We run our experiments using the Turkish Twitter dataset against ME classifier. Experimental evaluation shows that our IWD based feature selection method is able to select better features with respect to well-known ReliefF filter based feature selection method. The classification performance has been increased significantly over the baseline result that is shown in Table I. The F-score of the Twitter dataset is increased from 0.691 to 0.721 using IWD based feature selection method. The proposed method is effective in reducing the number of features so that it is suitable for classification of high dimensional data. By reducing the feature space, our system also reduces the time required to classify test dataset sharply without loss of accuracy in classification.

REFERENCES

[1] H. Liu, P. S. Yu, N. Agarwal, T. Suel, "Social computing in the blogosphere", *IEEE Internet Comput.*, vol. 14, no. 2, pp. 12–14, 2010. DOI: 10.1109/MIC.2010.39.

[2] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques", in *Empirical methods in natural language processing*, 2002, pp. 79–86. DOI: 10.3115/1118693.1118704.

[3] M. Hu, B. Liu, "Mining and summarizing customer reviews", in *10th ACM SIGKDD Int. Conf. Knowledge discovery and data mining*, 2004, pp. 168–177. DOI: 10.1145/1014052.1014073.

[4] A. Duric, F. Song, "Feature selection for sentiment analysis based on content and syntax models", *Decis. Support Sys.*, vol. 53, no. 4, pp. 704–711, 2012. DOI: 10.1016/j.dss.2012.05.023.

[5] A. Abbasi, H. Chen, A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", *ACM Trans. Inf. Syst*, vol. 26, no. 3, pp. 1–34, 2008. DOI: 10.1145/1361684.1361685.

[6] C. Nicholls, F. Song, "Comparison of feature selection methods for sentiment analysis", in *Proc. 23rd Canadian Conf. Advances in Artificial Intelligence (AI 2010)*, 2010, pp. 286–289. DOI: 10.1007/978-3-642-13059-5_30.

[7] B. Agarwal, N. Mittal, "Prominent feature extraction for review analysis: an empirical study", *J. Exp. Theor. Artif. Intell.*, vol. 28, no. 3, pp. 485–498, 2016. DOI: 10.1080/0952813X.2014.977830.

[8] G. Stylios, C. D. Katsis, D. Christodoulakis, "Using bio-inspired intelligence for web opinion mining", *Int. J. Comput. Appl.*, vol. 87, no. 5, pp. 975–8887, 2014. DOI: 10.5120/15207-3610.

[9] J. Kaur, S. S. Sehra, S. K. Sehra, "Sentiment analysis of Twitter data using hybrid method of support vector machine and ant colony optimization", *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 7, pp. 222–226, 2016.

[10] B. O. Alijla, L.-P. Wong, C. P. Lim, A. T. Khader, M. A. Al-Betar, "A modified Intelligent Water Drops algorithm and its application to optimization problems", *Expert Syst. Appl.*, vol. 41, no. 15, pp. 6555–6569, 2014. DOI: 10.1016/j.eswa.2014.05.010.

[11] L. A. Moncayo–Martinez, E. Mastrocinque, "A multi-objective intelligent water drop algorithm to minimise cost Of goods sold and time to market in logistics networks", *Expert Syst. Appl.*, vol. 64, pp. 455–466, 2016. DOI: 10.1016/j.eswa.2016.08.003.

[12] A. Vajdi, G. Zhang, Y. Wang, Y. Zhao, D. Liu, T. Wang, "A new approach based on intelligent water drops algorithm for node selection in service-oriented wireless sensor networks", in *IEEE Fourth Int. Conf. Big Data and Cloud Computing*, 2014, pp. 33–40. DOI: 10.1109/BDCloud.2014.45.

[13] T. Alexandros, D. Georgios, "Nature inspired optimization algorithms related to physical phenomena and laws of science: a survey", *Int. J. Artif. Intell. Tools*, vol. 26, no. 6, 2017. DOI: 10.1142/S0218213017500221.

[14] K. Kira, L. A. Rendell, "A practical approach to feature selection", in *Proc. Machine Learning*, 1992, pp. 249–256. DOI: 10.1016/B978-1-55860-247-2.50037-1.

[15] I. Kononenko, *Estimating attributes: Analysis and extensions of RELIEF*. Springer, Berlin, Heidelberg, 1994, pp. 171–182. DOI: 10.1007/3-540-57868-4_57.

[16] M. Robnik-Sikonja, I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF", *Mach. Learn.*, vol. 53, no. 1/2, pp. 23–69, 2003. DOI: 10.1023/A:1025667309714.

[17] E. Frank, M. A. Hall, I. H. Witten, *Data mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2016.

[18] E. Sarac, S. A. Ozel, "An ant colony optimization based feature selection for web page classification", *Scientific World Journal*, vol. 2014, 2014. DOI: 10.1155/2014/649260.

[19] E. Sarac, S. A. Ozel, "Web page classification using firefly optimization", in *2013 IEEE INISTA*, 2013, pp. 1–5. DOI: 10.1109/INISTA.2013.6577619.

[20] A. K. Uysal, S. Gunal, "A novel probabilistic feature selection method for text classification", *Knowledge-Based Systems,* vol. 36, pp. 226–235, 2012. DOI: 10.1016/ j.knosys.2012.06.005.

[21] C. D. Manning, H. Schutze, *Foundations of Statistical Natural Language Processing*. MIT Press Cambridge, 1999.

[22] M. Cetin, M. F. Amasyali, "Supervised and traditional term weighting methods for sentiment analysis", *21st Signal Process. Commun. Appl. Conf.*, pp. 1–4, 2013. DOI: 10.1109/SIU.2013.6531173.

[23] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006.

[24] E. Sarac, S. A. Ozel, "Effects of feature extraction and classification methods on cyberbully detection", *Suleyman Demirel Universitesi Fen Bilim. Enstitusu Derg.*, vol. 21, no. 1, p. 190, 2016. DOI: 10.19113/sdufbed.20964.

[25] T. Parlar, S. A. Ozel, "A new feature selection method for sentiment analysis of Turkish reviews", in *INISTA*, 2016, pp. 1–6. DOI: 10.1109/INISTA.2016.7571833.

[26] S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python*. O'Reilly Media. 2009.

[27] H. Shah-Hosseini, "Optimization with the nature-inspired intelligent water drops algorithm", in *Evolutionary Computation*, InTech, 2009. DOI: 10.5772/9623.