A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification

Seyyed Mohammad Hossein Dadgar / Master᾽s Student Artifical Intelligence
Department of computer Engineering
Islamic Azad University, Central Branch
Tehran, Iran
dadgar_1370@yahoo.com

Mohammad Shirzad Araghi / Master᾽s Student Artifical Intelligence
Department of computer Engineering
Islamic Azad University, Central Branch
Tehran, Iran
moh.shirzadaraghi.eng@iauctb.ac.ir

Morteza Mastery Farahani / Master᾽s Student Artifical Intelligence
Department of computer Engineering
Islamic Azad University, Central Branch
Tehran, Iran
farahani7431@yahoo.com

***Abstract*—With the development of weblogs and social networks, many news providers share their news headlines on different websites and weblogs. One of the main text mining topics is how to classify news into different groups. This study aims to classify news into various groups so that users can identify the most popular news group in the desired country at any given time. Based on Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM), a news classification method was proposed. The proposed approach is comprised of three different steps: 1) text preprocessing, 2) feature extraction based on TF-IDF, and 3) classification based on SVM. The proposed approach was evaluated using two BBC datasets and five groups of 20Newsgroup datasets. The classification precisions were obtained as 97.84% and 94.93% for BBC and 20Newsgroup datasets respectively. These are very desirable results in comparison with other classification methods.**

Keywords—Preprocessing, Feature Extraction, TF-IDF, Support Vector Machine, Dataset

I. INTRODUCTION (*HEADING 1*)

Text classification methods have drawn much attention in recent years and have been widely used in many programs. These techniques are essential because the textual data is swiftly rising with the passage of time. Text mining tools are required to perform indexing and retrieval of this rapidly growing text data. Text mining is the finding of some information which is previously unknown by extracting that information from large sets of unstructured text. Today, this un-structured data is growing as mostly the information is available in an electronic form such as e-mails, on World Wide Web, electronic publications and other documents. The term un-structured mean, the type of data in which the text is occurring in a natural free form or a sequence that may include word and sentence ambiguity. This un-structured information cannot be used for further processing by computers. The computers typically handle text as simple sequences of character string and are unable to provide useful information from the given text, without any process performed on it. Therefore, specific processing and preprocessing methods are required in order to extract useful patterns and information from the unstructured text [1].

During the last decade, the majority of important newspapers and magazines developed websites to present news and other materials. Reading important and interesting news is useful to users, but it is also time-consuming since they would have to read all the news items. Therefore, a news classification method for receiving the relevant information quickly seems to be essential. Many researchers have allocated much work to the automation of news classification in order to develop such a text classification system. Titles of several news classification methods proposed in the literature include : Financial News Classification [2], Classification of Short Texts [3], Automatic News Headlines Classification [4], a Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-Nearest Neighbor and Support Vector Machine [5], Classification of News Headlines for Providing User-Centered E-Newspaper [6], Emotions Extraction from News Headlines [7] ,and Short News Headlines Classification of Twitter [8].

Latest Research indicate that computer-based news classification is more efficient than classification by human beings. According to these studies, it takes a human being 89 hours to do the classification. Even if humans can classify more quickly, they can work only 8 hours a day,

978-1-4673-9916-6/16/$31.00 ©2016 IEEE

whereas computers can work 24 hours a day [9]. Therefore, a computer-based classification would be a desirable solution.

Section 2 describes the relevant researches on the subject. The proposed approach is introduced in Section 3 and evaluated in Section 4. Section 5 is devoted to discussion and conclusion.

II. PREVIOUS STUDIES

During recent years, many studies have been conducted on classification of news texts. Before introducing and investigating the proposed approach, we take a brief look at the methods previously presented in this regard.

In this paper, TF-IDF algorithm was used to classify news articles in Bahasa Indonesia. This algorithm counts the weight of each word with respect to its repetition in the text and the number of files in which it exists. When a word is repeated too many times in all the texts, it means that that word is not important, and that a high precision has been achieved in classification [9].

The Bayesian algorithm is a simple and efficient method used in text classification. However, it is not highly efficient because it does not model texts well, nor does it provide a good feature selection. In addition, there are other problems associated with this algorithm. This paper made some modifications to the Bayesian algorithm in order to improve its efficiency. Finally, the algorithm was used to group together the filtered spam messages [15].

A new supervised queue selection method for developing the similarity between a word and a class was introduced to improve the efficiency of text classification. In this method, each word is attributed a score based on its similarity to each class. Then the words are rated in terms of these scores [10].

Given the increasing number of studies on human emotion perception, an automated classifier system used the support vector machine to classify such human emotions as anger, hatred, fear, and grief in the word net affect dataset. The results indicated that SVM would lead to the best classification for emotion recognition in sentences [7].

Due to the popularity of the Internet, the number of online newspapers is increasing day by day, a fact which has made it difficult for users to access their favorite articles in digital newspapers. This paper proposed a news customization system based on SVM to recommend favorite articles to users with respect to their previously defined interests created in their profiles [6].

This paper implemented a classifier which operates by combining the nearest neighbor algorithm and the SVM classifier. This new hybrid method was named SVM-NN. The nearest neighbor algorithm was introduced as one of the most widely-used text classification algorithms due to its simplicity and efficiency in classifying different types of text. However, determining the effective K parameter to increase the algorithm precision remained a problem. In this paper, a hybrid SVM-KK method was proposed in order to minimize the effect of parameters on classification precision. In the training phase, the SVM was used to reduce the number of training samples in each group to its support vector. The support vectors from different groups were employed as training data in the nearest neighbor algorithm in which the Euclidean distance function was used to calculate the average distance between the test data and the support vectors in each group. Classification decisions were made based on the group wherein the shortest average distance between test data and support vectors existed. The tests indicated that classification precision of SVM-KK had a smaller effect on the values of parameters in comparison with KNN [5].

This comparison introduced a smart system for online news classification based on the hidden Markov model and the SVM. Considering the predefined classifications, a smart system was designed to extract the key words from the content of an online newspaper for the purpose of classification [11].

III. THE PROPOSED APPROACH

In this part, a brief explanation is presented on the software used for designing the proposed method. Then the proposed news text classification approach is investigated. This approach is comprised of three steps: text preprocessing, feature selection based on TF-IDF, and text classification using SVM. Each step is individually described in the following section. Figure 1 depicts the general architecture of the proposed method.

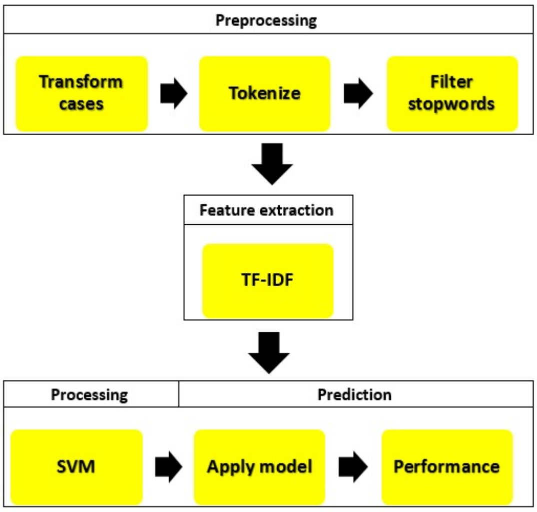

Figure 1.general architecture of the proposed method

A. Design Tools

It is highly important to select an appropriate software application because creating proper technical text mining conditions should include the three parameters of reliability, stability, and high computing speed. In this study, RapidMiner Studio Professional 6.5 was used to analyze data [13].

B. RapidMiner

RapidMiner is an open-source platform independently used for data mining. It is also employed as an integrated data mining engine with other products [14].In this platform, information is linked to operators and the information processed by each operator is transformed to the next operator in succession. Moreover, text preprocessing extension is used to preprocess text data [14].

C. Text Preprocessing

Text preprocessing is the first step in the process of news classification. A text is effectively preprocessed when unstructured data, mostly combinations of useful and useless data, are first received. The data are collected from different sources, and they should be cleaned. First, the text data are cleaned from distortion and useless information such as punctuations, exclamations, semicolons, irrelevant sentences, quotations, dates, etc.

D. Transforming Cases

The transform case operator transforms all the (upper case and lower case) characters existing in the text into lower case characters. This operator is used to eliminate homologous words which are different only in terms of their case.

E. Tokenizing

The majority of studies conducted on text mining include words or sentences which should be separated word by word for increased processing. Thus, all the words are separated in sentences, and all the punctuations are disposed of since they cannot represent any group. This simplifies computations in the next steps [9].

F. Filtering Stopwords

Filtering stopwordsis one of the methods used since the first studies were conducted on information retrieval. This algorithm is mainly employed to delete unnecessary things such as words occurring too frequently or too infrequently in sentences or text documents. It is also used to delete unimportant words or the words with no specific meanings such as *a, an,* or *the*. Like the previous steps, this step is applied in order to reduce processing time or computational complexity [9].

G. Feature Extraction based on TF-IDF

Upon creating the glossary, the weight of each word is calculated with respect to TF-IDF which is one of the most famous algorithms used in text mining research. The word *frequency* means the number of times a term is repeated in a text, and IDF stands for *Inverse Document Frequency*, an algorithm used to calculate the inverse probability of finding a word in a text [9].

Equation 1 is a classic TF-IDF equation used to calculate weight:

$$w_{ij} = tf_{ij} * \log \frac{N}{df_i} \quad (1)$$

In this equation, w_{ij} is the weight of weight of the word *i* in the document *j*, *N* is the number of documents in the set of total documents, tf_{ij} is the frequency of the word *i* in the document *j*, and df_i is the number of documents containing the word *i* [12].

H. Support Vector Machine

The support vector machine is a classification technique which was first applied by Joachims for text classification. This is a powerful and supervised learning sample based on the lowest structural risk principle. During training, this algorithm creates a hyper plane for separating positive and negative samples. Then it classifies new samples by specifying where on the hyper plane, each sample must be placed. In this study, nu-svc type is usedfor classification task. Furthermore, we chose rbf as its kernel type since we wanted to map our training samples nonlinearly to the higher dimensional space. Moreover, we set parameter, nu, to its maximum value 0.5. The majority of previous studies applying the SVM for text classification used all the words existing in text without considering the relevant importance of the words. On the other hand, various studies were conducted on keyword selection for text classification. As mentioned earlier, these studies were commonly focused on the criteria used for selecting keywords, using set- or keyword-class-based methods. Time complexity analysis and non-use of standard datasets created additional problems in these methods. Furthermore, the majority of studies did not use the SVM as the classifier algorithm [11].

IV. EVALUATION

In this section, the BBC and 20Newsgroup datasets are used to evaluate the proposed method. The explanation pertaining to the datasets and measurement criteria are given below.

A. Datasets

In this study, the 20Newsgroup dataset [16] and the BBC dataset [17] were used to evaluate the proposed method. The BBC dataset was provided by collecting information on five topics: business, entertainment, politics, sports, and technology, culminating in 2225 news texts from 2004 until 2005. The next dataset, which was used to evaluate the proposed method, was the 20Newsgroup dataset including 1997 news articles collected from the Internet in 20 classes. Only the subject and body of each text were used. Some of the text groups were so similar to each other (For example, texts on IBM hardware computer systems and Macintosh hardware computer systems). However, some other text groups were quite irrelevant (For example, text sets on sales and Christianity). Generally, the texts in this dataset were different from the previous texts because they included more words, and words would normally have more meanings. Furthermore, their writing style (dialog or email) is very different from other methods. In this study, 5070 news articles in five classes were used: computer graphics, objects offered for sale, as well as baseball, Christianity, and political texts on guns.

B. Performance Measures

In order to evaluate the effectiveness of class assignments, we use the standard recall, precision and f-measure aggregated over all classes. Precision and recall are defined as:

$$Recall = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ examples} \quad (2)$$

$$Precision = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ predictions} \quad (3)$$

The f-measure combines recall and precision with an equal weight in the following form:

$$F\text{-}measure = \frac{2*recall*precision}{recall+precision} \quad (4)$$

These scores are computed for the binary decisions on each individual class and then be averaged over all classes. A good algorithm should attain as high a recall value as possible without sacrificing precision. The closer the values of precision and recall, the higher is the f-measure. The value of f-measure lies between 0 and 1. A high f-measure value is desirable for good classification. Classification accuracy is also used for performance evaluation [10].

C. Results

Table 1 shows the precision of news texts classification in the two datasets (BBC and 20Newsgroup) obtained upon using the proposed method. The high precision of the proposed method is observed in the classification of news texts. According to this table, the classification precisions were 97.84% and 94.93% for the BBC dataset and the five evaluated groups in the 20Newsgroup dataset respectively. These levels of precision in text classification point to the high efficiency of this method. One of the reasons for this level of precision is that TF-IDF method was used for feature extraction. Table 2 shows the values of F obtained for the BBC datasets. The best group of this dataset, which could be classified, was sports with 99.22%, after which it was entertainment with 98.57%. The worst group was business with 96.70%. As explained in the previous section, the closer the value of F is to 1, the better the classification result shall be. According to this table, the majority of groups were close to 1, a fact which also indicated the high efficiency of the proposed method. Table 3 shows the values of F obtained for five classes of the 20Newsgroup dataset in which the best group was politics with 97.34% while the worst group was graphics with 91.65%.

TableI- precision of news texts classification

Method	**BBC (%)**	**20 Newsgroup (%)**
Our approach	97.84	94.93

TableII- values of F obtained for the BBC datasets

Group name	**F-measure**
Business	0.9670
Politics	0.9732
Entertainment	0.9857
Sport	0.9922
Tech	0.9736

TableIII- values of F obtained for five classes of the 20Newsgroup dataset

Group name	**F-measure**
Graphics	0.9165
Forsale	0.9279
Sport.baseball	0.9704
Religion.christian	0.9607
Politics.guns	0.9734

V. CONCLUSION

Given the high dimensions of the data involved, text classification is a challenging task. In this study, an approach was proposed to classify news texts. This approach was comprised of three different steps: 1) text preprocessing, 2) feature extraction based on TF-IDF, and 3) classification based on SVM. This approach was trained through the SVM classifier which was selected because it could support data with high dimensions. Using the BBC dataset and five groups of the 20Newsgroup dataset, the news group was evaluated. The classification precision was obtained 97.48% and 94.93% for the BBC dataset and the 20Newsgroup dataset respectively. These results were in good agreement with those obtained from other methods of text classification. Moreover, the F scale was used to obtain a single value between precision and recall. Through using this scale, we found the best group in the BBC set to be sports with 99.22%, whereas in the 20Newsgroup, the best group was politics with 97.34%.

REFERENCES

[1] M. I. Rana, S. Khalid, and M. U. Akbar, "News classification based on their headlines: A review," in IEEE 17th International Multi-Topic Conference (INMIC), 2014, pp. 211-216.

[2] B. Drury, L. Torgo, and J. J. Almeida, "Classifying news stories to estimate the direction of a stock market index," in 6th Conference on Information Systems and Technologies Iberian (CISTI), 2011, pp. 1-4.

[3] A. Heß, P. Dopichaj, and C. Maaß, "Multi-value classification of very short texts," in proceedings of the 31st annual German conference on Advances in Artificial Intelligence, Springer- Verlag Berlin, 2008, pp.70-77.

[4] M. W. Pope, "Automatic classification of online news headlines," University of North Carolina at Chapel Hill, november 2007.

[5] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," Expert Systems with Applications, 2012, pp. 11880-11888.

[6] R. Deshmukh and M. D. Kirange, "Classifying news headlines for providing user centered e-newspaper using SVM," in International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) , 2013, vol 2, Issue 3.

[7] D. Kirange, "Emotion classification of news headlines using svm," Asian Journal of Computer Science and Information Technology, 2012, pp. 104-106.

[8] I. Dilrukshi, K. De Zoysa, and A. Caldera, "Twitter news classification using SVM," in 8th International Conference on Computer Science & Education (ICCSE), 2013, pp. 287-291.

[9] A. A. Hakim, A. Erwin, K. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa

Indonesia based on term frequency inverse document frequency (TF-IDF) approach," in 6th International Conference on Information Technology and Electrical Engineering (ICITEE), 2014, pp. 1-4.

[10] T. Basu and C. Murthy, "Effective text classification by a supervised feature selection approach," in IEEE 12th International Conference on Data Mining Workshops (ICDMW), 2012, pp. 918-925.

[11] G. Krishnalal, S. B. Rengarajan, and K. Srinivasagan, "A new text mining approach based on HMM-SVM for web news classification," in International Journal of Computer Applications, , 2010, vol. 1, pp. 98-104.

[12] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of (TF-IDF), LSI and multi-words for text classification," Expert Systems with Applications, 2011, vol. 38, pp. 2758-2765.

[13] L. Drelichowski and J. Siwiec, "Application of text-mining for analysis and knowledge clustering," in scientific journal studies and proceedings of the Polish Association for Knowledge Management, 2012, No. 58.

[14] B. Wilges, R. C. Bastos, G. P. Mateus, and M. A. R. Dantas, "A case-comparison study of automatic document classification utilizing both serial and parallel approaches," Journal of Physics: Conference Series, 2014, vol. 540, p. 012001.

[15] G. Qiang, "An effective algorithm for improving the performance of naive Bayes for text classification," in Second International Conference on Computer Research and Development, 2010, pp. 699-701.

[16] data set 20 newsgroups collected by Ken Lang &homepage is http://www.ai.mit.edu/people/jrennie/20Newsgroups, Accessed dec 5 , 2015.

[17] data set BBC news collected by BBC news website & homepage is http://mlg.ucd.ie/datasets/bbc.html, Accessed dec 8, 2015.