Dokuz Eylül Üniversitesi-Mühendislik Fakültesi
Fen ve Mühendislik Dergisi
Cilt 19, Sayı 57, Eylül 2017

Dokuz Eylul University-Faculty of Engineering
Journal of Science and Engineering
Volume 19, Issue 57, September 2017

DOI: 10.21205/ deufmd.2017195762

Classification of Short-Texts by Utilizing an External Knowledge Source

Mert CALISAN[1], C. Okan SAKAR[*1]

Bahçeşehir Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, 34353, İstanbul

(Alınış / Received: 01.11.2016, Kabul / Accepted: 19.04.2017,
Online Yayınlanma / Published Online: 20.09.2017)

Keywords
Short-Text Classification, Enriched Text Representation, Bag-of-Words, Social Media

Abstract: The traditional text representation methods have been successfully applied to normal-length documents in many applications. However, these methods may worsen the performance of classifiers when used to represent short-texts in short-text classification tasks especially when the training set size is small. In this paper, we propose a method which is based on generating new feature representations of short-text by utilizing a knowledge base and using these representations together with the traditional feature representation method in classification problem. The experimental results show that using the proposed and traditional representations together improves the overall accuracy of the classifier.

Kısa-Metinlerin Harici Bir Bilgi Kaynağından Faydalanılarak Sınıflandırılması

Anahtar Kelimeler
Kısa Metin Sınıflandırma, Zenginleştirilmiş Metin Temsili, Kelime-Torbası, Sosyal Medya

Özet: Geleneksel metin temsil yöntemleri birçok uygulamada normal uzunluktaki dökümanlara başarıyla uygulanmıştır. Ancak bu yöntemler kısa-metinlerin sınıflandırılması problemlerinde kullanıldığında, özellikle eğitim veri kümesinin az sayıda örnekten oluşması durumunda sınıflandırıcının başarısını düşürebilmektedir. Bu makalede, bir bilgi tabanından faydalanarak kısa metinlerin yeni öznitelik temsillerinin oluşturulması ve geleneksel öznitelik temsil yöntemleri ile beraber sınıflandırma probleminde kullanılmasına dayanan bir yöntem önerilmiştir. Deney sonuçları önerilen ve geleneksel temsil yöntemlerinin birlikte kullanıldığında genel sınıflandırma başarısını artırdığını göstermektedir.

*Sorumlu yazar: okan.sakar@eng.bau.edu.tr

1. Introduction

Automated text classification is a popular research area in recent years due to increase in the volume of digital text produced in Internet media. The traditional Bag-of-Words (BoW) [1] feature representation method used to represent normal-length documents is based on using the frequency of each word as a feature to train a classifier. Considering that some words appear more frequently in general and may not contain discriminative information, to improve the BoW representation more sophisticated numerical statistics such as term frequency–inverse document frequency (tf-idf) [2] have been proposed in the literature, in which the frequency of a word in a document is offset by the frequency of the word in the corpus.

The BoW representation has been successfully applied to normal-length documents with proper natural language processing techniques. However, using BoW method to represent short-texts which contain only a few words results in very sparse matrices which significantly worsens the performance of the classifiers [3, 4]. Considering the increasing usage of content sharing services, social networking platforms, news portals, and e-commerce sites, the development of feature representation techniques focusing on enrichment of limited number of words in a short-text is needed to cope with the sparsity problem of data matrices obtained with BoW representation. Another problem encountered when BoW is used to represent short-texts is the loss of semantic information [5]. In the BoW method, each word is independently represented by disregarding the word order and grammar. Although this problem can be tolerated in the presence of long documents, short-texts suffer from loss of semantic information more since they contain very brief information consisting of a few words.

The proposed methods that have been addressed in the literature to overcome the limitations of BoW in short-text classification are based on using different representations or extending BoW representation. Utilisation of external knowledge, which is taken from an external data source, is a widely-used approach for short-text classification. The enrichment of the data using external knowledge addresses both sparseness and lack of semantic information problems [5]. As an external knowledge, using web-based encyclopedia improves the performance short-texts classification systems especially when there is limited number of samples [6]. For example, Wikipedia [7] domain specific features like article titles, anchor texts, and figure captions possess important semantic value that can be used to enrich the content of short-text documents [8]. Other than domain specific features, raw text can also be used and mined for term association rules to enrich the short-text instances [9].

To the best of our knowledge, there is only one study that used Turkish Wikipedia (Vikipedi) to enhance document classification in Turkish language and it addresses normal-length documents [10]. Specifically, they utilized the titles (concepts) of Vikipedi articles to enrich sample documents consisting of Turkish newspaper articles. If a title is found in sample document, it is added to the vector representation of it as a single term even title is a multiple word title. Four different representations are produced and the classification accuracies obtained using these representations are compared. These representations are BoW, BoW enriched with Multiple World Vikipedi Concepts, samples represented only with found Multiple Word Vikipedi Concepts and samples represented with mapped both

Single and Multi Word Vikipedi Concepts. They applied Naïve Bayes Multinomial (NB) and Support Vector Machine (SVM) classifiers individually for each representation, and showed that augmenting traditional BoW representation with Vikipedi concepts improves the accuracy of classifiers. In this paper, we propose a feature representation method based on Vikipedi knowledge base to overcome the limitations of BoW representation when used for short-text classification with small-sample sized dataset. The representations generated using knowledge base include both domain specific properties of Vikipedi and n-gram based semantic features. The proposed representations are not considered as alternatives of BoW representation but as additional sources of information which can improve the accuracy and generalization of BoW representation when used together. Therefore, in addition to these representations, traditional BoW representation has also been generated and all representations have been combined in an ensemble manner [11] to improve the generalization and accuracy of the final model. The experimental results show that combining the predictions of the proposed representation methods with those of the BoW representation significantly improves the success of the model especially when the number of training samples is limited.

2. Material and Method

The dataset used in this study were extracted from TS Abstract Corpus [12] and a Turkish News data [13]. There are 6 different classes in the dataset and an instance is composed of 146 characters and 18 tokens in average. A knowledge base, which covers the classes in seed short-text dataset, is created from Vikipedi. As the first step of knowledge base creation, we identified one Vikipedi article category for each class label. Then, we selected articles in these categories according to their relevance to create a knowledge base with high quality. In the tree-like structure of Vikipedi, articles are leafs and categories are nodes, where an article can have multiple direct parent nodes. We observed that increase in the depth between a category and descendent article decreases the relevance of article content in accordance with category. This is because the number of direct or indirect parent nodes increases with depth and nested tree-like structure. As an example, Biology is indirect category parent of Ice Cream article where the path follows as Biology Category, Biological Energy Category, Food Category, Meals Category and Ice Cream article. Due to this reason, we selected articles which are direct or second level descendent of identified categories. Besides, we have filtered the articles which belong to multiple identified categories/classes. Thus, we exported the building blocks of our knowledge base consisting of 2490 selected articles in Extensible Markup Language (XML) format that Vikipedi provides. Then, we processed the articles in XML form and extracted both the plain article texts and domain specific properties of Vikipedi. These domain specific properties include descriptions of images in the articles, subtitles, link texts to other articles, and many others which have semantic value. All extracted texts were tokenized, lemmatized and stored with a weight value assigned to each token. By this way, tokens along with the category information representing their class labels, weight values representing their frequencies and tag information indicating if they are extracted from a domain specific property or not are obtained, which constitutes the knowledge base used in this study. For tokenization and lemmatization natural language

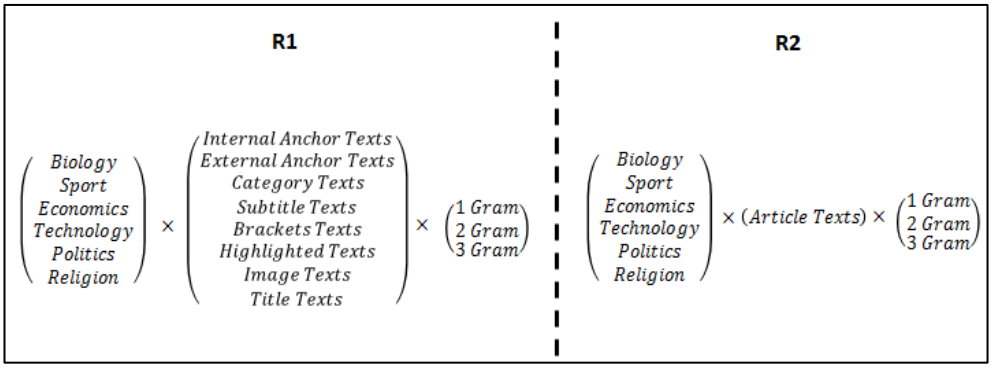


Figure 1. Feature combinations of the short-texts generated from knowledge base: R1 and R2 Features

processing (NLP) tasks, an open source Turkish NLP library called Zemberek [14] was used while a variation of tf-idf scheme was applied to weight the frequencies of terms.

After creating the knowledge base, we generated multiple representations of seed short-text dataset to be used for classification. We generated three different representations of seed short-text: Vikipedi domain specific representation as R1, Vikipedi article plain text representation as R2, and traditional BoW representation with tf-idf weighting as R3. In R1 and R2 representations, we define one feature for each Class–Element–N-Gram combination, where class represents each category considered in this study, element is a domain specific property (e.g. article title in R1, and article plain text in R2). Thus, we represent a seed short-text with Vikipedi articles' domain specific properties in R1 and Vikipedi article plain texts in R2. For instance, Biology–Title–2-Gram combination of a seed short text in R1 represents the frequency of 2-gram tokens in this short-text in the titles of Vikipedi articles which are in biology category. There are 144 features for R1 and 18 features for R2 generated based on this approach. In Figure 1, all possible feature combinations of R1 and R2 are given.

To represent a short-text in R1 and R2 space [15], we first tokenize and lemmatize each term in the seed short-text. Then, we search tokens as 1-gram, 2-gram- and 3-gram in the knowledge base for their identified weight values specific to Class–Element–N-Gram combination. Sum of the found tokens' weight values for a Class–Element–N-Gram combination corresponds to the value of the feature. Thus, we populate feature values of a short-text instance by applying this method for each combination. An illustration of this scheme is given in Figure 2.

Our features generated for both R1 and R2 utilize 2-gram and 3-gram tokens, which increases the amount of semantic information preserved in the representations. Besides, R1 contains additional semantic information with domain specific properties. R2 has plain article text based features which are not specific to Vikipedi and can be generated from any plain text source providing that it includes category information of the source articles.

In R3 representation of seed short-text dataset, we use traditional BoW approach by lemmatizing each term and

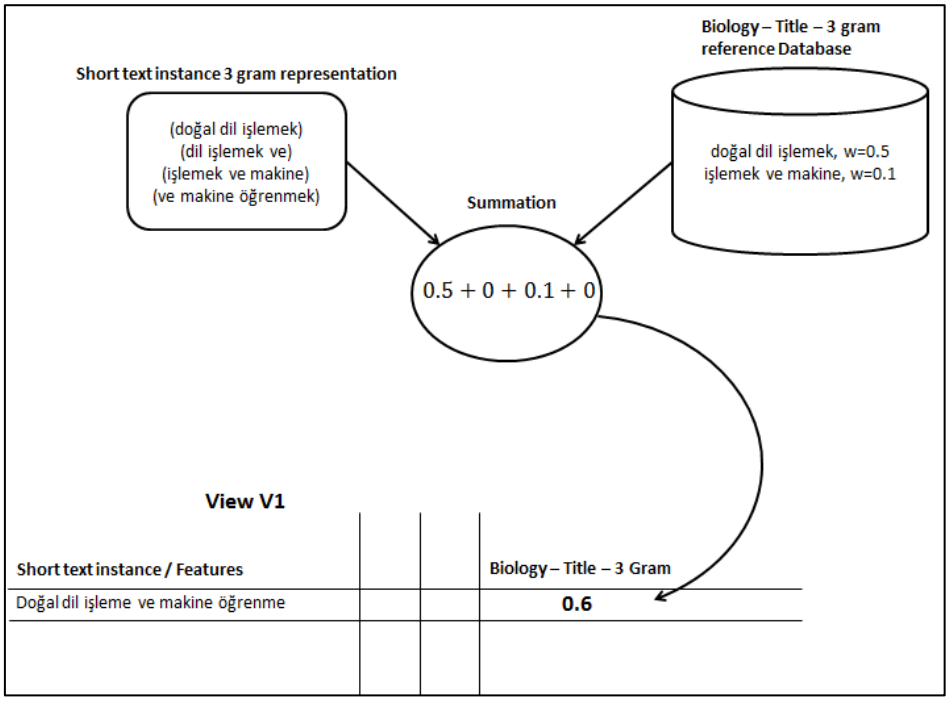


Figure 2. Populating feature values

assigning them a weight using tf-idf weighting scheme without using knowledge base. This results in a high dimensional feature representation when compared to R1 and R2.

We used our short-text dataset with multiple representations in a supervised learning manner. In our method, we first train an individual supervised classifier for each representation of short-texts, and apply each model to the test set instances. Then, we combine the predictions of the models to obtain the final decision of the ensemble. We used LIBSVM [16] implementation of Support Vector Machines (SVM) as the classifier, and combined the class posterior probability estimates of the network members using simple voting.

3. Results

As a pre-processing step, before feeding R1 and R2 representations to SVM classifier, we eliminated the features with zero variance. We observed that 29 features out of 144 have zero variance in R1 representation. Specifically, 5 of these features were 2-gram combinations while 24 of them were 3-gram combinations. These results show that although domain specific features of Vikipedi have adequate coverage when used to enrich normal-length documents, they fail to provide extra information when used for short-texts especially about 3-gram terms, which results in loss of semantic information. We should note that all of the R2 features have non-zero variance. Table 1 shows the number of instances per class in our dataset.

Table 1. Dataset size showing number of instances per class.

Class	# of instance
Biology	147
Sport	181
Economics	207
Technology	182
Politics	155
Religion	120
Total # of instances	992

In our experiments, we split the short-text dataset into train, test, and validation sets. We used training set to train SVM, validation set to optimize the SVM parameters, and test set to report the final accuracies obtained with the optimal parameters determined on the validation set. All experiments were repeated 10 times with random dataset splits for statistical significance and average accuracies of these 10 trials are reported. We performed paired t-test to evaluate the performance of the individual representations and their ensembles. We repeat this procedure with various numbers of training set instances while keeping validation set size constant as 20 instances per class and using the rest of the data for test set.

Table 2 shows the accuracies of individual short-text representations and their ensembles. The obtained results show that the highest accuracy was achieved with BoW representation (R3) for all training set sizes. Both R1 and R2 performed significantly worse than BoW (p-value < 0.05). However, in line with the general aim of this study, combining the predictions of R1 and R2 with those of the BoW representation improved the accuracy of BoW representation significantly when a limited number of labelled training instances are available (20 and 40 instances per class). We should also note that the superiority of the ensemble model to BoW diminishes with increasing number of training instances.

Table 2. SVM classification accuracies (%) obtained with individual representations and their ensembles

# of training instances per class	R1	R2	R3	Ensemble
20	62.6	62.6	72.5	78.3
40	69.4	67.1	82.3	85.6
60	70.6	67.0	85.4	86.7
80	73.7	71.0	87.8	89.0

3. Discussion and Conclusion

In this paper, a classification method based on multiple representations of short-text is proposed for short-text classification problem with small amount of training data. Our method utilizes a knowledge base, which is created from Vikipedi, to generate two new representations of short-text. Compared to traditional BoW representation, the proposed representations do not lead to sparse data matrices and contain semantic features. The obtained results show that SVM models obtained with R1 and R2 achieve comparable performances with limited number of training instances per class. However, with increasing number of training instances, SVM with domain specific features of the knowledge base performs better than that of the article plain text based features. This shows that domain specific features contain important precise information which can be helpful in classifying short-text documents. Besides, it was found that

combining the predictions obtained with the proposed knowledge-based representation and traditional Bag-of-Words representation in an ensemble manner results in a model with improved predictive performance when there is a small amount of training data. This makes the generated representations useful especially for semi-supervised learning tasks where the number of labeled samples is small compared to the number of unlabeled samples and multiple representations are required to incorporate the unlabeled training instances to the training phase.

References

[1] Gupta, V., Lehal, G.S. 2009. A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, Vol. 1(1), pp. 60-76.

[2] Salton, G., McGill, M.J. 1986. Introduction to Modern Information Retrieval, New York: McGraw-Hill, Inc., 440p.

[3] Man, Y., 2014. Feature Extension for Short Text Categorization Using Frequent Term Sets. Procedia Computer Science, Vol. 31, pp. 663-670.

[4] Wang, B.K., Huang, Y.F., Yang, W.X. and Li, X., 2012. Short Text Classification based on Strong Feature Thesaurus. Journal of Zhejiang University-Science C, Vol. 13(9), pp. 649-659.

[5] Huang, A., Milne, D., Frank, E. and Witten, I.H. 2008. Clustering Documents with Active Learning Using Wikipedia. Eighth IEEE International Conference on Data Mining, December 15-19, Pisa, 839-844.

[6] Gabrilovich, E., Markovitch, S. 2006. Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge. National Conference on Artificial Intelligence (AAAI), AAAI Press, July 16-17, Boston, 1301-1306.

[7] Milne, D., Witten, I.H. 2008. Learning to Link with Wikipedia. 17th ACM conference on Information and knowledge management, October 26-30, CA, 509-518.

[8] Zhang, Z., Lin, H., Li, P., Wang, H., Lu, D. 2013. Improving Semi-Supervised Text Classification by Using Wikipedia Knowledge. International Conference on Web-Age Information Management, June 14-16, Beidaihe, China, 25-36.

[9] Man, Y. 2014. Feature Extension for Short-Text Categorization Using Frequent Term Sets, Procedia Computer Science, Vol. 31, pp.663-670

[10] Poyraz, M., Ganiz, M.C., Akyokuş, S., Görener, B., Kilimci, Z.H. 2012. Exploiting Turkish Wikipedia as a Semantic Resource for Text Classification. International Symposium on Innovations in Intelligent Systems and Applications (INISTA), Trabzon, 1-5

[11] Dietterich, T.G. 2002. Ensemble Learning, In the Handbook of Brain Theory and Neural Networks, Second edition, (M.A. Arbib, Ed.), Cambridge, MA: The MIT Press, pp.405-408

[12] Ozturk, S., Sankur, B., Gungor, T., Yilmaz, M. B., Koroglu, B., Agin, O.,

… & Ahat, M. 2014. Turkish Labeled Text Corpus. 22nd Signal Processing and Communications Applications Conference (SIU), April 23-25, Trabzon, 1395-1398.

[13] Amasyalı, M.F., Yıldırım, T. 2004. Otomatik Haber Metinleri Sınıflandırma. Signal Processing and Communications Applications Conference (SIU), April 28-30, Aydın, 224-226.

[14] Akın, A.A., Akın, M.D., 2007. Zemberek, An Open Source NLP Framework for Turkic Languages. Structure, Vol. 10, pp.1-5.

[15] Calisan, M. 2015. Multi-View Short-Text Classification Using Knowledge Bases. Bahçeşehir University, Graduate School of Natural and Applied Sciences, Master Thesis, 50p, İstanbul.

[16] Chang, C.C. Lin, C.J., 2011. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 2(3), p.27.