Short text classification using deep representation: A case study of Spanish tweets in Coset Shared Task

Erfaneh Gharavi and Kayvan Bijari

Faculty of New Science and Technologies,
University of Tehran, Tehran, Iran
{e.gharavi,kayvan.bijari}@ut.ac.ir

Abstract. Topic identification as a specific case of text classification is one of the primary steps toward knowledge extraction from the raw textual data. In such tasks, words are dealt with as a set of features. Due to high dimensionality and sparseness of feature vector result from traditional feature selection methods, most of the proposed text classification methods for this purpose lack performance and accuracy. In dealing with tweets which are limited in the number of words the aforementioned problems are reflected more than ever. In order to alleviate such issues, we have proposed a new topic identification method for Spanish tweets based on the deep representation of Spanish words. In the proposed method, words are represented as multi-dimensional vectors, in other words, words are replaced with their equivalent vectors which are calculated based on some transformation of raw text data. Average aggregation technique is used to transform the word vectors into tweet representation. Our model is trained based on deep vectorized representation of the tweets and an ensemble of different classifiers is used for Spanish tweet classification. The best result obtained by a fully connected multi-layer neural network with three hidden layers. The experimental results demonstrate the feasibility and scalability of the proposed method.

Keywords: Deep representation; Word Vector Representation; Spanish Tweet Classification.

1 Introduction

Topic identification is one of the primary steps toward text understanding. This process has to be done automatically due to the large amount of existing texts. A few number of topic identification applications are as follow [18].

- **News filtering and organizing**
- **Opinion Mining**: Mining people's opinion on social events, like an election, in order to predict the trend toward such events.
- **Email Classification and Spam Filtering**: Automatic classification of emails is desirable in order to determine the subject or to find out junk email [1].

People use social medias to state their idea about social events. People's tweets capture researchers' attention on different issues and they have been used for a variety of purposes, such as marketing communication, education and politics [3]. Specifically about the political conventions such as election which make individuals active and leads to thousands of tweets. Topic identification of a given tweet is the first step toward its analysis.

There are a lot of challenges regarding classifying tweets includes colloquialism in tweets, spelling variation, use of special characters, violating regular grammar rules etc [2]. The short text does not provide sufficient word occurrences.

In order to work with textual data, it should be described numerically to enable computers to process that. In traditional approaches, words are considered as distinct features for representing textual data. Those representations suffer from scarcity and disability to detect synonyms [5]. To avoid these issues which required time-consuming feature engineering, deep learning techniques are used which proves its competency in many application such as NLP [8]. The essential goal of deep learning [13] is to improve all procedures of NLP in an efficient way. Deep representation of text data makes it easy to compare words and sentences as well as minimizing the need to use lexicons.

In this paper, we used deep text representation to classify Spanish tweets. We applied the method provided in our previous work [10] to this classification Shared task to assess the feasibility of our approach in variety of natural language processing task in different languages.

The structure of this paper is as follows: first, related works are studied in Section 2, deep representation of text is described in Section 3. An introduction to tweet classification using deep representation is given in Section 4. Section 5 deals with an introduction to the evaluation metric, and results of the proposed approach over the provided data sets. Finally, Section 6 ends the paper with conclusion and some insights.

2 Related Works

In this section, some well-known methods for text classification are described and specifically some of the recently presented tweet classification approaches are discussed and reviewed.

In the work of Ghavidel et al.[11] after selecting keywords using standard term frequency/inverse document frequency (TF-IDF), vector space method was applied for classifying a Persian text corpus consisting 10 categories. Li et al [16] classify text by combining the functionality of Support Vector Machine (SVM) and K-nearest neighbor for text classification, precision of 94.11(%) was resulted for 12 topics. In [6], BINA et al. used K-nearest neighbor algorithm using statistical features including 3-grams, 4-grams and three criteria including Manhattan, Dice and dot product were used for classifying Hamshahri corpus. As a feature for text classification, emoticons are used in [20] for tweet topic classification. Dilrukshi et al.[9] Classified Twitter feeds using SVM Text classifier. Bakliwal

et al.[2] used Stanford and Mejaj data set to find sentiments in its tweets. They assign a positive and a negative probability to each word, to find sentiments of each tweet. Wikipedia and Wordnet used in [14] to cluster short texts accurately. Malkani & Gillie[17] used supervised SVM, Neural Network (NN), Naive Bayes (NB), and Random Forest for Twitter topic classification. SVM outperforms other supervised classification algorithms. Zubiaga et al.[23] extracted 15 language independent features from the trending topics. They trained an SVM classifier with the features and it is then used to classify trending topics.

3 Deep Text Representation

Deep learning tries to find more abstract features using deep multiple layer graph. Each layer has linear or non-linear function to transform data into more abstract ones [4]. Hierarchical nature of concept makes new feature representation a suitable approach for natural language processing. Advantages of using deep methods for NLP task are listed below:

- No hand crafted feature engineering is required
- Fewer number of features in comparison to the traditional methods
- No labeled data is required [7].

3.1 Word Vector Representation

In application of deep representation in natural language processing, each word is described by the surrounding context. This vector which is generated by a deep neural networks and contain semantic and syntactic information about the word. In distributed word representation, generally known as word-embedding, the similar words have the similar vectors [21]. Skip-grams and continuous bag of words, which are employed by this study, are two-layer neural networks that are trained for language modeling task [7].

3.2 Text Document Vector Representation

A composition function should be provided for combining word vectors to represent text. Paragraph Vector is an unsupervised algorithm that used the idea of word vector training and considered a matrix for each piece of text. This matrix also update during language modeling task. Paragraph vector outperforms other methods such as bag-of-words models for many applications [15]. Socher [21] introduce Recursive Deep Learning methods which are variations and extensions of unsupervised and supervised recursive neural networks (RNNs). This method encodes two word vectors into one vector by auto-encoder networks. Socher also presents many variation of these deep combination functions such as Matrix-Vector Recursive Neural Networks (MV-RNN) [22]. There are also some simple mathematical methods which applied as a composition function generally used as benchmarks [19].

4 Tweet Classification using Deep Text Representation

In this section, we describe our approach for Spanish tweet classification. These steps include pre-processing that describe the text refinement process, then composition function to combine the word embedding to construct and represent each tweet is illustrated. Then the classification algorithms applied to classified the tweets into the aforementioned categories. Figure 1 shows the steps of tweet topic identification procedure.

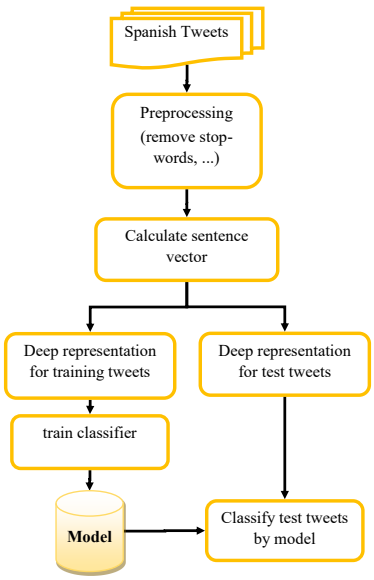

Fig. 1. Spanish tweets classification steps

4.1 Pre-processing

As the first step of the processing, the following steps are done as pre-processing of every block of text, i.e tweet. Such as elimination of special characters as: "&", "(", ")", "#", and removal of all numbers.

Other most common pre-processing functions is removing the stop-words. In this regard, we simply omit list of stop words available on-line[1]. This list includes 178 stop words like: "una", "es", "soy", "vamos", etc.

[1] http://www.ranks.nl/stopwords/spanish

4.2 Tweet Representation

We first retrieve Spanish 300-dimensional word vectors on-line[2]. Then Spanish stop words eliminated while text pre-processing. After that, for each sentence an average of all word vectors is calculated as in equation (1).

$$S_i = \frac{\sum_{i=1}^{n} w_i}{n} \tag{1}$$

Where S is the vector representation for tweet and w_i is the word vector for $i-th$ word of the sentences and n is the number of words in that sentence.

We represent each tweet in train and test corpus by this approach. For each tweet in training, validation and test corpus we have a 300-dimensional text representation. These vectors considered as feature vectors which are required to classify the tweets.

5 Experimental Evaluation

In this section, at first, dataset used in the Coset Shared task will be described. Then evaluation metrics for tweet classification are defined.

5.1 Dataset

In COSET Shared task, provided dataset consist of 5 issues: political issues, related to the most abstract electoral confrontation; policy issues, about sectoral policies; personal issues, on the life and activities of the candidates; campaign issues, related with the evolution of the campaign; and other issues. The tweets are written in Spanish and they talk about the 2015 Spanish General Election. The training set consists 2242 tweets. The development contains 250 tweets to help participants to train their model and test it on 624 test tweets[3].

5.2 Evaluation Criteria

The metric used for evaluating the participating systems was the F1 macro measure. This metric considers the precision and the recall of the systems predictions, combining them using the harmonic mean. Provided that the classes are not balanced, the Coset committee proposed using the macro-averaging method for preventing systems biased towards the most populated classes.

$$F1_{macro} = \frac{1}{L} \sum_{l \in L} F_1(y_1, \hat{y}_1) \tag{2}$$

[2] http://crscardellino.me/SBWCE/
[3] http://mediaflows.es/coset/

5.3 Results

The results of applying the proposed method over training Spanish Tweet corpus is presented in table 1, The final results as well as rankings are reported in [12].

Vectorized tweets are trained and tested via different learning algorithms such as random forest, support vector machine with linear kernel, naive bayse, and logistic regression. Table 1 shows the achieved results of different learning algorithms on the Spanish corpus.

Table 1. Result of model training over basic learning algorithms

Algorithm	Precision	Recall	F-Measure	Rank
SVM	0.61	0.62	**0.60**	1
Random Forest	0.59	0.58	0.55	2
Naive Bayse	0.61	0.46	0.48	3
Logestic Regression	0.57	0.54	0.49	4

Furthermore, since neural network structures usually grasp a better intuition of deeply extracted set of features from the datasets, vectorized sentences are feed into a multi layer perception neural network with three hidden layers inside. table 2 shows results of the neural network and its characteristic over the given dataset.

Table 2. Result of model training over MLP structures

N.N.(hidden neurons)	Precision	Recall	F-Measure	Rank
MLP (500; 240; 100)	0.63	0.64	**0.63**	1
MLP (600; 300; 150)	0.58	0.59	0.58	3
MLP (300; 150; 50)	0.61	0.61	0.60	2
MLP (50; 75; 25)	0.57	0.59	0.57	4

Based on the achieved experimental results of studied algorithms, it is clear that neural network structure is a splendid choice to deal with vectorized sentences for performing Twitter topic classification task. In this regard and with numbers of trials, a multi layer perceptron neural network with 3 hidden layers each containing 500, 240, and 100 neurons is selected for the Coset shared task.

6 Conclusion

In this paper, we proposed a method for topic identification as a typical case of text categorization on Spanish tweets by using the deep representation of words. Results from experimental evaluations showed the feasibility of our approach. With no hand engineering feature and simple composition function, we achieved 55(%) based on $F1_{macro}$ score. The best method reported 64(%) of $F1_{macro}$ [12].

In order to improve the proposed method, we can consider the following ideas: In this work, we used words which are included in Spanish word vectors, while it is clear that incorporating stemming would help us to find all words in the tweet on that list. We can also provide methods to deal with out of vocabulary words. In addition, considering the unknown topic in the proposed method is another issue in this area.

Furthermore, given its admirable runtime, our proposed method is scalable and is applicable for a large number of documents and can be used for practical purpose. As a future work, we are going to apply other composition functions and also try word-by-word vector comparison in order to eliminate drawbacks of the current method.

Acknowledgments

The authors would like to thank the reviewers for providing helpful comments and recommendations which improve the paper significantly.

References

1. Aggarwal, C., Zhai, C.: A Survey of Text Classification Algorithms. Mining Text Data pp. 163–222 (2012)
2. Bakliwal, A., Arora, P., Madhappan, S., Kapre, N.: Mining sentiments from tweets. Proceedings of the WASSA 12 (2012)
3. Batool, R., Khattak, A.M., Maqbool, J., Lee, S.: Precise tweet classification and sentiment analysis. In: 2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS). pp. 461–466. IEEE (jun 2013)
4. Bengio, Y.: Learning Deep Architectures for AI. Foundations and Trends® in Machine Learning 2(1), 1–127 (2009)
5. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A Neural Probabilistic Language Model. The Journal of Machine Learning Research 3, 1137–1155 (2003)
6. BINA, B., AHMADI, M. H. & RAHGOZAR, M.: Farsi Text Classification Using N-Grams and Knn Algorithm A Comparative Study. In: Data mining. pp. 385–39 (2008)
7. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. Proceedings of the 25th international conference on Machine learning pp. 160–167 (2008)
8. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research 12, 2493–2537 (2011)
9. Dilrukshi, I., De Zoysa, K., Caldera, A.: Twitter news classification using SVM. In: Proceedings of the 8th International Conference on Computer Science and Education, ICCSE 2013. pp. 287–291 (2013)
10. Gharavi, E., Bijari, K., Veisi, H., Zahirnia, K.: A Deep Learning Approach to Persian Plagiarism Detection. Working notes of FIRE 2016-Forum for Information Retrieval Evaluation (2016)
11. Ghavidel Abdi, H., Vazirnezhad B., Bahrani, M.: Persian text classification. In: 4th conference on Information and Knowledge Technology (2012)

12. Giménez, M., Baviera, T., Llorca, G., Gámir, J., Calvo, D., Rosso, P., Rangel, F.: Overview of the 1st Classification of Spanish Election Tweets Task at IberEval 2017. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Murcia, Spain, September 19, CEUR Workshop Proceedings. CEUR-WS.org, 2017.
13. Hinton, G.E., Osindero, S., Teh, Y.W.: A Fast Learning Algorithm for Deep Belief Nets. Neural Computation 18(7), 1527–1554 (2006)
14. Hu, X., Sun, N., Zhang, C., Chua, T.s.: Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge. Proceedings of the 18th ACM conference on Information and knowledge management pp. 919–928 (2009)
15. Le, Q.V., Mikolov, T.: Distributed Representations of Sentences and Documents 32 (2014)
16. Li, W., Miao, D., Wang, W.: Two-level hierarchical combination method for text classification. Expert Systems with Applications 38(3), 2030–2039 (2011)
17. Malkani, Z., Gillie, E.: Supervised Multi-Class Classification of Tweets (2012)
18. Margarida De, A., Cachopo, J.C., Bernard, D.P., Doutor, B., Emílio, J., Pavão, S., Doutora, M., Maria, I., Trancoso, M., Arlindo, D., Limede De Oliveira, M., Mário, D., Gaspar, J., Silva, D., Pável, D., Calado, P.: Improving Methods for Single-label Text Categorization (2007)
19. Mitchell, J., Lapata, M.: Composition in Distributional Models of Semantics. Cognitive Science 34(8), 1388–1429 (2010)
20. Read, J.: Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. ACL Student Research workshop (June), 43–48 (2005)
21. Socher, R.: Recursive Deep Learning for Natural Language Processing and Computer Vision. PhD thesis (August) (2014)
22. Socher, R., Huval, B., Manning, C.D., Ng, A.Y.: Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. pp. 1201–1211. Association for Computational Linguistics (2012)
23. Zubiaga, A., Spina, D., Martnez, R., Fresno, V.: Real-time classification of Twitter trends. Journal of the Association for Information Science and Technology 66(3), 462–473 (2015)