Ensembles of methods for Tweet Topic Classification

Gretel Liz De la Peña Sarracén

Center for Pattern Recognition and Data Mining, Cuba
gretel@cerpamid.co.cu,
http://www.cerpamid.co.cu

Abstract. This paper describes the system we developed for IberEval 2017 on Classification Of Spanish Election Tweets (COSET) task. Our approach is based on a weighted average ensemble of five classifiers: 1) a classifier based on logistic regression; 2) a support vector machine classifier; 3) a Naive Bayes classifier for multinomial models; 4) a Guassian Naive Bayes classifier; and 5) a classifier implementing the k-nearest neighbors vote. Each such classifier was choice taking into account its contributes to the success of the system. The aim is to design a approach by using a voting method, where individual classifiers can have weaknesses. The performance of the ensemble is compared to the individual classifiers, and the experimental results show that the ensemble has better results.

Keywords: Ensemble Classifier, Tweets Classification, Spanish Election Tweets, Twitter

1 Introduction

Nowaday, Twitter[1] have become an important part of the daily life of many of users. This microblogging services are used as communication media, recommendation services, real-time news sources and information sharing sites. The large amount of new data created as result makes automatic analysis essential for processing this data. Thus, Twitter has become an attractive area for many studies such as text classification.

Text classification aims at labeling natural language texts into a fixed number of predetermined categories [1–3]. On Twitter, users post short text messages called tweets which makes quite difficult this task because of their features. Tweets are small (only 140 characters) and are charectized by their informal style language, many grammatical errors and spelling mistakes, slang and vulgar vocabulary, and abbreviations.

The Classification Of Spanish Election Tweets (COSET) task, in IberEval 2017 workshop [4] has as main goal to classify a corpus of political tweets in 5 categories of classification: political issues, related to the most abstract electoral

[1] http://www.twitter.com

confrontation; policy issues, about sectorial policies; personal issues, on the life and activities of the candidates; campaign issues, related with the evolution of the campaign; and other issues. This paper presents an ensemble-based approach developed to participate in this task. The main objective is to explore and identify the advantages for designing a better approach by using a voting method which improves the performance of individuals classifiers.

The paper is organized as follows. Section 2 describes our system. Next, in Section 3, the experimental results are discussed. Finally, we present our conclusions in Section 4 with a summary of our findings.

2 System

Our system carries out a sequence of steps, which goes from text preprocessing to tweet classification. The classification approach, based on the comination of five classifier in order to assign to a tweet a class label, is described below.

2.1 Preprocessing

In the preprocessing step, tweets are cleaned. First, taking advantage of regular expressions, the emoticons are detected and removed from the text. Also, we eliminate all links, urls and user names which can be identified because their first character is the symbol @. As regards the words starting with hashtags (that is, the symbol #), we do not realize modifications because they can be related directly with the topic of the text. Finally, we convert the words to lowercase and remove all non-letters characters and all stopwords present in tweets.

2.2 Methods

Classifier based on Logistic Regression (LR): Logistic regression belongs to the family of classifiers known as the exponential or log-linear classifiers. It works by extracting some set of weighted features from the input, combining them linearly, and then applying an exponential function to this combination. Thus, Logistic regression is a discriminative model that assigns a class to an observation by computing a probability from the function of a weighted set of features of the observation [5, 6].

Logistic Regression is good at dealing with very high dimension data. Text classification is a classic problem.

Support Vector Machine classifier (SVM): Support Vector Machine uses linear models to implement nonlinear class boundaries. It transforms the input space using a nonlinear mapping into a new space. Then a linear model constructed in the new space can represent a nonlinear decision boundary in the original space. It plots each data item as a point in n-dimensional space (where

n is number of features) with the value of each feature being the value of a particular coordinate. Then, it performs classification by finding the hyper-plane that differentiate the classes very well.

Support Vector Machine classifier works really well with clear margin of separation and is effective in cases where number of dimensions is greater than the number of samples [7].

Guassian Naive Bayes classifier (GaussianNB): Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes theorem with the assumption of independence between every pair of features. In spite of their apparently over-simplified assumptions, naive Bayes classifiers have worked quite well in many real-world situations. They require a small amount of training data to estimate the necessary parameters.

Naive Bayes classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality [8]. In Guassian Naive Bayes classifier the likelihood of the features is assumed to be Gaussian.

Naive Bayes classifier for Multinomial Models (MultinomialNB): Multinomial Naive Bayes is used for multinomially distributed data. It is one of the classic naive Bayes variants most used in text classification.

Classifier based on k-nearest neighbors vote (KNN): Neighbors-based classification is a type of instance-based learning or non-generalizing learning. This is it does not attempt to construct a general internal model, but simply stores instances of the training data. Classification is computed from a simple majority vote of the nearest neighbors of each point, where a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. The number of nearest neighbors taken into account can be a constant k. In our case, we take $k = 1$ for its contributes to the success of the system. Despite its simplicity, it is often successful in classification situations where the decision boundary is very irregular [9].

Ensemble classifier (Ensemble): Our design of classifier is an ensemble of the above classifiers. In this way, the predicted class probabilities for each classifier are collected, multiplied by the classifier weight, and averaged. The final class label is then derived from the class label with the highest average probability. We used scikit-learn with these purposes [10].

We find the best setting of weights via brute force grid search, limiting the coefficient values in the interval [0,1], and taking into account the performance of base algorithms. Those with better results were assigned higher weights.

3 Results

In order to evaluate the advantages of the ensemble-based approach, we have used the F1 macro measure. This metric considers the precision and the recall of the systems predictions, combining them using the harmonic mean. Specifically, we rely on the macro for preventing systems biased towards the most populated classes. Table 1 shows the results for the ensemble classifier we have proposed as well as for the base classifiers on the validation set. As expected, according to the measure, the ensemble classifier reveals a marked performance improvement with the highest score.

Table 1. Performance on the validation set

Classifier	$F_{1-macro}$	Classifier	$F_{1-macro}$
LR	0.5377	SVM	0.5423
MultinomialNB	0.4510	GaussianNB	0.5093
KNN	0.4693	Ensemble	**0.5847**

Further, we have studied the impact of other text processing beyond the preprocessing described above. Thus, once the set of simple rules of preprocesing has been applied, we have realized other preprocesses. We have assigned, to each word in the text, its lemma and have tried with some feature selection methods.

Feature selection The main objective of feature selection methods is to decrease of the dimensionality of the dataset by eliminating features that are not related for the classification [11]. We have tried with two methods. On one side, we have removed features with low variance (*VarSelection*), specifically all zero-variance features. On the other hand, we have applied univariate feature elimination (*UFSSelection*), which works by selecting the 10 best features based on univariate statistical tests, the χ^2 test in our case.

Table 2. Performance of ensemble approach with different preprocessing

Method	$F_{1-macro}$	$F_{1-macro}$ (lemmatized texts)
No feature selection	0.5847	0.6254
VarSelection	0.5818	**0.6273**
UFSSelection	0.5818	0.6197

Table 2 shows that the results do not improve with the feature selection methods. However, with the lemmatized texts, a remarkable improvement in the performance of ensemble approach is achieved, even for the cases where the feature selection is applied, obtaining the best results for the method that removing features with low variance, although the results were similar.

4 Conclusion

This paper has described an ensemble-based approach for IberEval 2017 on Spanish Election Tweets Classification task. We combined five classiers by a weighted average, with the aim of designing an approach which improves the performance of the base classifiers. The results showed that, indeed, the ensemble classifier reveals a remarkable performance improvement with the highest score. Also, as part of experiments, we studied some preprocessing for texts: lemmatization and features selection methods. We achieved the best results with lemmatized texts and the feature selection method that removing features with low variance.

References

1. Durga Bhavani Dasari et al. Text categorization and machine learning methods: current state of the art. *Global Journal of Computer Science and Technology*, 12(11-C), 2012.
2. Vandana Korde and C Namrata Mahender. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2):85, 2012.
3. Rajni Jindal, Ruchika Malhotra, and Abha Jain. Techniques for text classification: Literature review and current trends. *Webology*, 12(2):1, 2015.
4. Gimnez M., Baviera T, Rosso P., Llorca G., Gmir J., Calvo D., and 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.*
5. Raymond E Wright. Logistic regression. 1995.
6. Alexander Genkin, David D Lewis, and David Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.
7. Thorsten Joachims. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.
8. Haiyi Zhang and Di Li. Naïve bayes text classifier. In *Granular Computing, 2007. GRC 2007. IEEE International Conference on*, pages 708–708. IEEE, 2007.
9. Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers. *Multiple Classifier Systems*, 34:1–17, 2007.
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
11. George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305, 2003.