Dealing with High Dimensional Sentiment Data Using Gradient Boosting Machines

Vasileios Athanasiou(✉) and Manolis Maragoudakis

Department of Information and Communication Systems Engineering,
University of the Aegean, Palama 2 Street, 83200 Karlovasi, Samos, Greece
{icsdml5041,mmarag}@aegean.gr

Abstract. One of the most common classification tasks that applies on textual information is sentiment analysis, i.e. the prediction of the sentiment of a given document. With the vast use of social media and internet applications such as e-commerce, e-tourism and e-government, numerous comments and opinions are broadcasted per day, thus an automatic way of analyzing them is of great importance. The present paper focuses on sentiment analysis for Greek texts, obtained from Web 2.0 platforms. Greek is a language that lacks an in-depth availability of natural language processing tools in the sense that most of them are not publicly available. The novelty of the article is that instead of utilizing preprocessing tools such as Part-of-Speech taggers, text stemmers and polar-word lexica, it incorporates the translation of the Greek token as provided by the Google Translator® API. Since automatic translation of Greek sentences often results in poor translations where the meaning of the original sentence is severely deteriorated, the translation of each token individually is almost 100 % correct. However, taking the translation of every Greek token poses a significant issue to the outcome of the classification process for practically any classifier, therefore, we introduce the use of a powerful ensemble algorithm that is highly customizable to the particular needs of the application, such as being learned with respect to different loss functions and thus dealing with a large number of dimensions. This algorithm is called Gradient Boosting Machines and experimental results support our claim that it surpasses other, well-known machine learning techniques with a significant improvement for our task.

Keywords: Gradient Boosting Machines · Sentiment analysis · High-dimensional data · Modern Greek

1 Introduction

Throughout recent years, we have witnessed a remarkable rise of the Internet and the World Wide Web that has altered the way they communicate, seek for relevant information, and work on the most fundamental level. The arrival of Web 2.0 technologies has resulted in a vast popularity and ubiquity of web resources built around the ideas of social media and user-generated content (e.g., Facebook, Twitter, and LinkedIn). In the beginning such concepts were part of more traditional web resources such as online retailers (e.g., Amazon) or electronic media, in order to having their products or articles enhanced with the users' opinions. Therefore, the task of relevant

© IFIP International Federation for Information Processing 2016
Published by Springer International Publishing Switzerland 2016. All Rights Reserved
L. Iliadis and I. Maglogiannis (Eds.): AIAI 2016, IFIP AICT 475, pp. 481–489, 2016.
DOI: 10.1007/978-3-319-44944-9_42

information from the vast amounts of human communication information over the Internet is of utmost importance for robust sentiment analysis modules. In fact, the origin of opinionated data has caused the development of Web Opinion Mining (WOM) [1], as a new concept in Web Intelligence. WOM deals with the issue of extracting, analyzing and aggregating web data about opinions. The analysis of users' opinions is significant because through them it is possible to determine how people feel about a product or service and know how it was received by the market. In general, traditional sentiment analysis mining techniques apply to social media content as well, however, there are certain factors that make Web 2.0 data more complicated and difficult to be parsed. An interesting study about the identification of such factors was made by Maynard et al. [2], in which they exposed important features that pose certain difficulties to traditional approaches when dealing with social media streams. Modern Greek language is posing additional obstacles to sentiment analysis since the majority of preprocessing tools for Greek such as POS tagger, stemmer and polarity lexica are not freely available.

In the present paper, we deal with modeling a sentiment analyzer for Modern Greek based on a simple, yet novel idea. We bypass the need for extensive preprocessing tools and utilize a freely available translation API that is provided by Google®, in order to augment the feature set of the training data. However, since automatic translation of sentences often suffer from poor performance, mainly due to the large degree of ambiguity, we decided to translate each Greek token individually, a process that rarely makes mistakes. The resulted feature set was of course double the original size which also poses certain difficulties to the majority of classification algorithms. Hence, we experimented with an ensemble classification algorithm named as Gradient Boosting Machines (GBM) which is theoretically proven of being able to cope with large number of features [3]. According to the referenced study, "a possible explanation why boosting performs well in the presence of high-dimensional data is that it does variable selection (assuming the base learner does variable selection) and it assigns variable amount of degrees of freedom to the selected predictor variables or terms". We experimented with numerous well-known algorithms using the initial Greek-only feature set as well as the enhanced translated one and also some basic feature reduction techniques such as Principal Component Analysis [4] and found that GBM are superior to any other implementation.

2 Related Work

The nature as well as the popularity of feedback-oriented content in domains such as news, social events, services and products is something of a duty to the human urge to post what they feel and think online. As soon as the most common type of message on Twitter is about 'me now' [5], it is evident that users talk often about their own feelings and opinions. Bollen et al. [6] stated that users express both their own mood in tweets about themselves and more generally in messages about other subjects. Another study [7] estimates that approximately 19 % of microblog messages mention a brand name and from those that do, around 20 % contain sentiment.

There are numerous challenges in applying typical sentiment analysis and opinion mining techniques to social media. Short texts in social media, known as micro-posts, are, perhaps, the most challenging text type for opinion mining, given the fact that they do not contain much contextual information and take much implicit knowledge into account. Ambiguity is a common phenomenon since one cannot easily make use of co-reference information: unlike the situations of blog posts and comments, tweets do not typically follow a conversation thread, and appear much more in isolation from other tweets. They also contain much more language variation, tend to be less structured than longer posts, contain unorthodox forms of writing such as emoticons, abbreviations and hashtags, which can form an important part of the meaning. Typically, they contain extensive use of irony and sarcasm, which are particularly difficult for a machine to identify. On the contrary, their shortness can also be beneficial in focusing the topics more explicitly: it is very uncommon for a single tweet to be related to more than one topic, which can therefore contribute in disambiguation.

The research of [8] presents a wide-ranging and detailed review of traditional automatic sentiment detection techniques, including many sub-components, which we shall not repeat here. In general, sentiment detection techniques can be roughly divided into lexicon-based methods (e.g. [9]) and machine learning methods, e.g. [10]. Lexicon-based methods rely on a sentiment lexicon, a collection of known and pre-compiled sentiment terms. Machine learning approaches make use of shallow syntactic and/or linguistic features [11, 12], and hybrid approaches are also very common, with sentiment lexicons playing a key role in the majority of methods, e.g. [13]. Decision Trees, Naïve Bayes and SVM (Support Vector Machine) have been applied from the supervised side [13, 14]. On the unsupervised side there is a pattern-logic classification according to a lexicon. Combination of the above can characterized as semi-supervised [13]. In "On Mining Opinions From Social Media" [15] the authors confirm that Naïve Bayes outperforms a large number of other methods. Techniques that work based on rules normally use sentiment dictionaries plus rules [16].

3 Gradient Boosting Machines

Generally, the boosting methods work by adding sequentially new models and in every iteration every weak base-learner is re-trained [3]. The target is to reduce the total error of the model. The GBM works by constructing new base-learners which are correlated with the negative gradient of the loss function. The loss function can be set from the user (Gaussian L2 loss function, Binomial etc.), this is the main reason that GBM is highly customizable. Suppose there is a dataset $(x,y)_{i=1}^{N}$ where $x = (x_1, x_2 ... x_d)$ are the input variables and y the corresponding labels. The relation between x and y is unknown, this means it is needed to reconstruct this relation in a way to minimize a loss function.

$$\hat{f}(x) = y, \hat{f}(x) = arg \min_{f(x)} \Psi(y, f(x))$$

Where $f(x)$ indicates the dependence between x, y, $\hat{f}(x)$ is the estimation function and $\Psi(y,f(x))$ is the loss function. The same problem expressed in terms of expectations concludes to

$$\hat{f}(x) = arg \min_{f(x)} \underbrace{\underbrace{E_x[E_y(\Psi[y,f(x])|x]}_{expectation\ over\ the\ whole\ dataset}}_{expected\ loss}$$

where $E_y(\Psi[y,f(x])$ is the response variable depended from x. Significant notation is that y can became from not only one distribution provides the ability for more loss functions Ψ.

In machine learning, there are cases that the data set produce a vector with very high dimensionality, for example in sentiment analysis. That means the rows of the document term matrix have very low density because the appearance of the words in every observation is very rare. GBM can to overcome this difficulty because is able to create sparse models, additionally can work with different types of base learners. The ability of use different loss functions combining the ability to use different type of learners provides great regularization capabilities. This has impact to avoid overfitting and achieve generalization to our model. Apart from the above advantages GBM algorithm has drawbacks such as memory consumption and evaluation speed. The first one has impact to the other. GBM works by doing iterations which stored in memory, easily in hard cases for the algorithm like intrusion detection systems the number if iteration can be tens of thousands. This problem appears to all ensemble algorithms.

4 Experimental Setup and Results

4.1 Data Collection

The corpus of data retrieved from online newspaper articles. The types of articles came from financial domain, political, society and sports in plain text type and Greek language. All articles have been selected under the condition to be able classified as positive or negative only. The corpus is balanced and two datasets have been derived from it. The first consists from the articles as gathered in Greek language (Single Dataset), the second one from the article in Greeks plus the translation in English (Mixed Dataset).

4.2 Data Preprocessing

The data included URLs, which have been automatically removed using an HTML tag identifier. As mentioned before, due to lack of availability of Greek preprocessing tools, only some base linguistic tools have been incorporated. The preprocessing phase consists of the following steps:

- Tokenization: To extract only the words, all stop-words have been removed, all letters lowercased.

- Translation of each Greek token using the Google® Translation API.
- Creation of a document-term matrix that contains n rows, where n equals the number of comments (n was 520 for our case, having 180 positive and 340 negative ones) and m rows where m corresponds to the size of the Greek vocabulary, almost double when applied the translation, since some tokens shared the same translation. In each cell of the document-term table, the value of the tf-idf weight of each term is contained, given by the following formula:

  $tf - idf(term) = frequency(term) \cdot log\left(\frac{m}{N(term)} + 1\right)$, where m is the total number of terms in the collection and N (term) is a function that returns the number of documents the term appears in.
- Dimensionality reduction using the Principal Component Analysis (PCA): This step applied only in cases we used Naïve Bayes, Decision Trees and Support Vector Machines (SVM) and not in GBM since it could cope with the large number of attributes.

PCA reduces the dimensionality by performing transformation of possibly correlated variables to a fully new dataset with linearly uncorrelated variables, called principal components. Every principal component has variance starting from the first one which has the largest value and this means the item has the biggest impact in dataset comparing the other principal components. All the principal components are the eigenvectors of the covariance matrix, that means are orthogonal.

4.3 Evaluation Criteria and Performance

For the experiments have been used three measures accuracy, precision and recall as described below:

$$\text{Accuracy: } \frac{T_p + T_n}{T_p + T_n + F_p + F_n}, \text{ Precision: } \frac{T_p}{T_p + F_p}, \text{ Recall: } \frac{T_p}{T_p + F_n},$$

where T_p, T_n (true positive, true negative) are the correct positive and correct negative predictions respectively and F_p, F_n (false positive, false negative) are the wrongly positive and wrongly negative predictions respectively. In order to evaluate the performance for every classifier we used the 10-fold cross validation approach. Figure 1 illustrates the methodology in 3 sub-processes. The first two are preprocess the data. The first one produce the Single dataset. The second one produce the translated dataset, adding the Single dataset with translated dataset produces the Mixed dataset. The third sub-process applies the classifier to the input dataset.

4.4 Classification Benchmark

In order to examine the performance of the proposed GBM method and evaluate it against other, well-known classifiers that have previously applied in sentiment analysis tasks with success, as explained in the related work section, the following classifiers

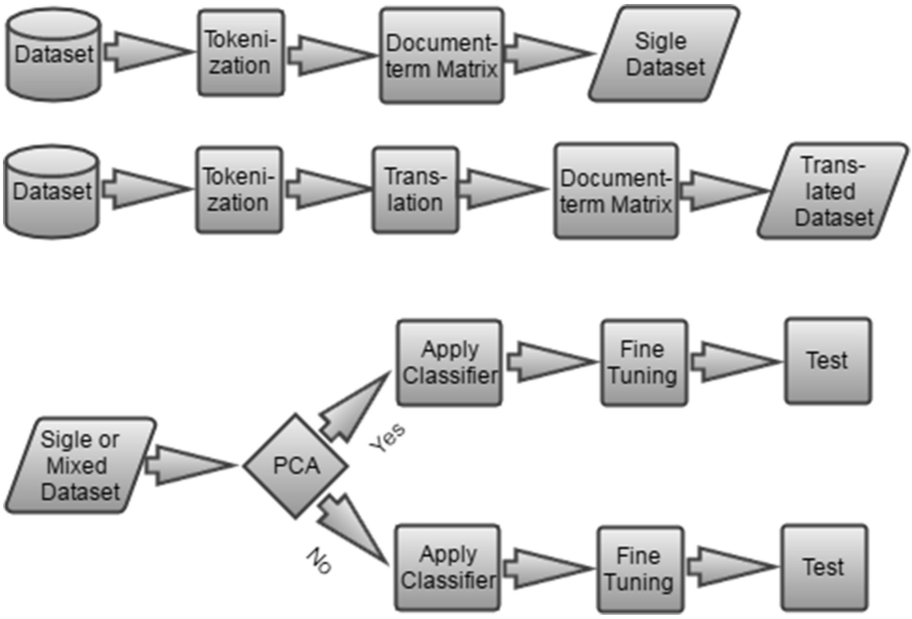

Fig. 1. Methodology flowchart

that have been additionally incorporated: Decision Trees, Naïve Bayes, SVM, Gradient Boosting Machine and Deep Neural Network Learning. In the case of Decision Trees, Naïve Bayes and SVM we have experimented using two different data representation techniques, namely the original tf-idf vector representation of the document-term matrix and the reduced dataset of the PCA transformation using the first 5 principal components.

4.5 Experimental Results

As previously explained, we incorporated two different datasets based on the same set of annotated documents. The first set, called as the **Single** dataset contained only the Greek tokens. The second set, called as the **Mixed** set, contained both Greek tokens and their translation. Finally each of the aforementioned datasets was also undertaken the PCA data dimensionality technique. For reasons of space, we provide detailed outcomes for the Mixed dataset and discuss what happens when considering the other option. As regards to the Decision Tree classifier without performing PCA technique in the Mixed Dataset the outcome is depicted on Table 1.

Table 1. Decision Trees performance on the Mixed Dataset.

Accuracy 60.62 %	Positive	Negative
Precision	42.00 %	64.07 %
Recall	17.80 %	85.64 %

When applied PCA on data the performance has been decreased by 10 %. In addition, when Decision Trees tried on the Single dataset the performance was decreased by 6 %. The reason of this is that the Mixed dataset is has more examples thus our model trained better. For Naïve Bayes classifier without performing PCA on the Mixed Dataset the outcome is depicted on Table 2.

Table 2. Naïve Bayes performance on the Mixed dataset.

Accuracy 81.25 %	Positive	Negative
Precision	82.80 %	81.94 %
Recall	65.25 %	92.08 %

When applied PCA on data the performance in accuracy has been decreased more than 35 %. In addition when Naïve Bayes tried on the Single dataset the performance was decreased by 5 %. For SVM classifier without performing PCA technique on the Mixed dataset the outcome is depicted on Table 3.

Table 3. SVM performance on the Mixed dataset

Accuracy 71.11 %	Positive	Negative
Precision	60.15 %	78.84 %
Recall	66.67 %	73.76 %

When applied PCA on data the performance in accuracy has been decreased by 20 %. In addition when SVM tried on the Single dataset the performance was decreased by 5 %. For the GBM classifier in the Mixed Dataset the outcome is depicted on Table 4.

Table 4. GBM performance on the Mixed dataset

Accuracy 87.26 %	Positive	Negative
Precision	83.19 %	89.66 %
Recall	82.50 %	90.10 %

In Single dataset the performance of GBM in accuracy is 7 % lower. For Deep Learning classifier in Mixed Dataset the outcome is depicted on Table 5.

Table 5. Deep Learning performance on the Mixed dataset

Accuracy 74.69 %	Positive	Negative
Precision	62.34 %	85.88 %
Recall	80.00 %	71.57 %

In Single dataset the performance in accuracy of Deep Learning is 5 % lower. The above tables shows that GBM classifier has the best accuracy, precision and recall comparing with the all of other classifiers we have used. The unique exception is in Naïve Bayes classifier recall in negatives in which the Naïve Bayes is 1 % higher. The comparison of GBM and Naïve Bayes it seems clear in the next diagrams which Illustrates the performance of GBM, Naïve Bayes and Naïve Bayes after preprocessing the corpus with PCA. These diagrams are separated by the performance in Positives on the left and Negatives observations on the right (Fig. 2).

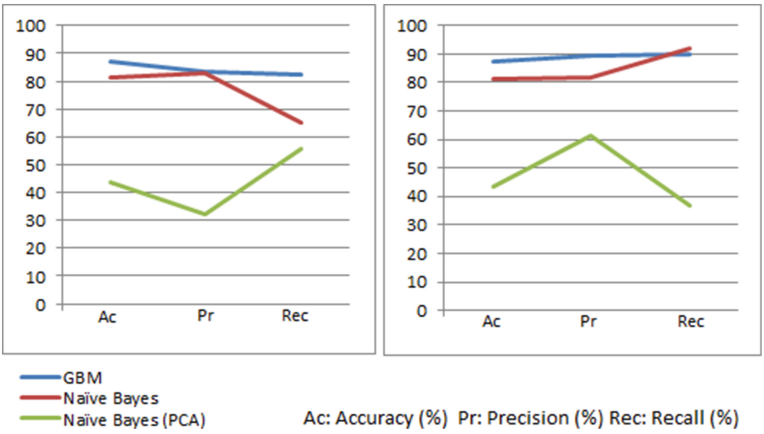

Fig. 2. Performance for GBM, Naïve Bayes and Naïve Bayes with PCA

In both of the cases GBM method slightly takes precedence over Naïve Bayes. In addition GBM algorithm seems to have stable in performance in both cases positives and Negatives on the Recall measure. That means Naïve Bayes in false negatives performed differently. The GBA classifier performs well even in cases we have two languages the Mixed Dataset. The difference is 7 %, in case the Naïve Bayes classifier the difference is at the same level 6 %.

5 Conclusions

This work dealt with the issue of performing sentiment analysis on Greek texts, obtained from Web 2.0 platforms using limited NLP resources. The novelty lied on the fact that instead of utilizing a manual polarity lexicon for Greek, which would have inadequate impact on social media linguistic style, we proposed a method that takes the translation of each token as additional input features. Even though this process may seem to bear additional effort and complexity to the majority of classification algorithms, the use of GBM, a robust boosting approach that can cope with high dimensional data appeared to be beneficial for the task at hand, outperforming a family of

well-known methods for sentiment analysis. In the future we can try this methodology on imbalanced dataset to see the performance. On this benchmark we can include more classifiers like Recurrent Neural Networks.

References

1. Taylor, E.M., Rodríguez, O.C., Velásquez, J.D., Ghosh, G., Banerjee, S.: Web opinion mining and sentimental analysis. In: Velásquez, J.D., Palade, V., Jain, L.C. (eds.) Advanced Techniques in Web Intelligence-2. SCI, vol. 452, pp. 105–126. Springer, Heidelberg (2013)
2. Maynard, D., Bontcheva, K., Rout, D.: Challenges in developing opinion mining tools for social media. In: Proceedings of @NLP can u tag #usergeneratedcontent? Workshop at LREC 2012, Turkey (2012)
3. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Front. Neurorobotics **7**, 21 (2011)
4. Shlens, J.: A tutorial on principal component analysis, derivation, discussion and singular value decomposition, 25 March 2003
5. Naaman, M., Boase, J., Lai, C.: Is it really about me? message content in social awareness streams. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, pp. 189–192. ACM (2010)
6. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: twitter sentiment and socio-economic phenomena (2009). http://arxiv.org/abs/0911.1583
7. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: tweets as electronic word of mouth. JASIST **60**(11), 2169–2188 (2009)
8. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retrieval **2**(1–2), 1–135 (2008)
9. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **1**, 1–41 (2011)
10. Boiy, E., Moens, M.-F.: A machine learning approach to sentiment analysis in multilingual web texts. Inf. Retrieval **12**(5), 526–558 (2009)
11. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Technical report CS224N Project Report, Stanford University (2009)
12. Pak, A., Paroubek, P.: Twitter based system: using twitter for disambiguating sentiment ambiguous adjectives. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 436–439 (2010)
13. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis
14. Bengio, Y.: Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009)
15. Politopoulou, V., Maragoudakis, M.: On mining opinions from social media. In: Iliadis, L., Papadopoulos, H., Jayne, C. (eds.) EANN 2013, Part I. CCIS, vol. 383, pp. 474–484. Springer, Heidelberg (2013)
16. Maynard, D., Funk, A.: Automatic detection of political opinions in tweets. In: García-Castro, R., Fensel, D., Antoniou, G. (eds.) ESWC 2011. LNCS, vol. 7117, pp. 88–99. Springer, Heidelberg (2012)