algorithms

MDPI

Article

A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources Are Not Plentiful: A Case Study for Modern Greek †

Vasileios Athanasiou and Manolis Maragoudakis *

Artificial Intelligence Laboratory, University of the Aegean, 2 Palama Street, 83200 Samos, Greece; icsdm15041@aegean.gr

* Correspondence: mmarag@aegean.gr; Tel.: +30-22730-79571

† This paper is an extended version of our paper "Dealing with High Dimensional Sentiment Data Using Gradient Boosting Machines" presented In Proceedings of the 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, 16–18 September 2016.

Academic Editors: Katia Lida Kermanidis, Christos Makris, Phivos Mylonas and Spyros Sioutas
Received: 11 December 2016; Accepted: 24 February 2017; Published: 6 March 2017

Abstract: Sentiment analysis has played a primary role in text classification. It is an undoubted fact that some years ago, textual information was spreading in manageable rates; however, nowadays, such information has overcome even the most ambiguous expectations and constantly grows within seconds. It is therefore quite complex to cope with the vast amount of textual data particularly if we also take the incremental production speed into account. Social media, e-commerce, news articles, comments and opinions are broadcasted on a daily basis. A rational solution, in order to handle the abundance of data, would be to build automated information processing systems, for analyzing and extracting meaningful patterns from text. The present paper focuses on sentiment analysis applied in Greek texts. Thus far, there is no wide availability of natural language processing tools for Modern Greek. Hence, a thorough analysis of Greek, from the lexical to the syntactical level, is difficult to perform. This paper attempts a different approach, based on the proven capabilities of gradient boosting, a well-known technique for dealing with high-dimensional data. The main rationale is that since English has dominated the area of preprocessing tools and there are also quite reliable translation services, we could exploit them to transform Greek tokens into English, thus assuring the precision of the translation, since the translation of large texts is not always reliable and meaningful. The new feature set of English tokens is augmented with the original set of Greek, consequently producing a high dimensional dataset that poses certain difficulties for any traditional classifier. Accordingly, we apply gradient boosting machines, an ensemble algorithm that can learn with different loss functions providing the ability to work efficiently with high dimensional data. Moreover, for the task at hand, we deal with a class imbalance issues since the distribution of sentiments in real-world applications often displays issues of inequality. For example, in political forums or electronic discussions about immigration or religion, negative comments overwhelm the positive ones. The class imbalance problem was confronted using a hybrid technique that performs a variation of under-sampling the majority class and over-sampling the minority class, respectively. Experimental results, considering different settings, such as translation of tokens against translation of sentences, consideration of limited Greek text preprocessing and omission of the translation phase, demonstrated that the proposed gradient boosting framework can effectively cope with both high-dimensional and imbalanced datasets and performs significantly better than a plethora of traditional machine learning classification approaches in terms of precision and recall measures.

Keywords: gradient boosting machines; sentiment analysis; high-dimensional data; Modern Greek

1. Introduction

Undoubtedly, the way that machines and humans search, retrieve and manage information changes at a high rate. The potential information resources have increased because of the advent of Web 2.0 technologies. User-generated content is available from a large pool of sources, such as online newspapers, blogs, e-commerce sites, social media, etc. The quantity of information is continuously increasing in every domain. Initially, it was online retailers, such as Amazon, that identified possible profits by exploiting users' opinions. However, nowadays, almost everyone has realized that quality of services, marketing and maximization of sales cannot be achieved without considering the textual content that is generated by Internet users. Therefore, the task of identifying relevant information from the vast amount of human communication information over the Internet is of utmost importance for robust sentiment analysis. In fact, the existence of opinion data has resulted in the development of Web Opinion Mining (WOM) [1], as a new concept in web intelligence. WOM focuses on extracting, analyzing and combining web data about user thoughts. The analysis of users' opinions is substantial because they provide the elements that determine how people feel about a topic of interest 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 parse. 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. The key difference lies in the fact that users are not passive information consumers, but are also prolific content creators. Social media can be categorized on a diverse spectrum, based on the type of connection amongst users, how information is shared and how users interact with the media streams:

- Interest-graph media [3], such as Twitter, encourage users to form connections with others, based on shared interests, regardless of whether they know the other person in real life. Connections do not always need to be reciprocated. Shared information comes in the form of a stream of messages in reverse chronological order.
- Professional Networking Services (PNS), such as LinkedIn, aim to provide an introduction service in the context of work, where connecting to a person implies that you vouch for that person to a certain extent and would recommend them as a work contact for others. Typically, professional information is shared, and PNS tend to attract older professionals [4].
- Content sharing and discussion services, such as blogs, video sharing (e.g., YouTube, Vimeo), slide sharing (e.g., SlideShare) and user discussion forums (e.g., CNET). Blogs usually contain longer contributions. Readers might comment on these contributions, and some blog sites create a time stream of blog articles for followers to read. Many blog sites also advertise automatically new blog posts through their users' Facebook and Twitter accounts.

The following challenging features are also recognized by researchers and classified as openings for the development of new semantic technology and text mining approaches, which will be better suited to social media streams:

- Short messages (microtexts): Twitter and most Facebook messages are very short (140 characters for tweets). Many semantic-based methods reviewed below supplement these with extra information and context coming from embedded URLs and hashtags. For instance, in the work of [5], the authors augment tweets by linking them to contemporaneous news articles, whereas in [6], online hashtag glossaries are exploited to augment tweets.
- Noisy content: social media content often has unusual spelling (e.g., 2moro), irregular capitalization (e.g., all capital or all lowercase letters), emoticons (e.g., :-P) and idiosyncratic abbreviations (e.g., ROFL for "Rolling On Floor Laughing", ZOMG for "Zombies Oh My God"). Spelling and capitalization normalization methods have been developed [7], coupled with studies

of location-based linguistic variations in shortening styles in microtexts [8]. Emoticons are used as strong sentiment indicators in opinion mining algorithms.

The Modern Greek language poses additional complications to sentiment analysis since the majority of preprocessing tools for Greek, such as the Part-of-Speech (POS) tagger, shallow syntactic parsers and polarity lexica, are not freely available. In the current work, we deal with modeling a sentiment analysis framework for Modern Greek, based on a simple, yet functional and effective idea. We propose an alternative approach that capitalizes on the power of existing induction techniques while enriching the language of representation, namely exploring new feature spaces.

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. Nevertheless, since automatic translation of large sentences is often reported as portraying low accuracy, mainly due to the large degree of ambiguity, especially in the case of Modern Greek, we followed a simple approach, translating each Greek token individually, a process that rarely makes errors. The resulting feature set portrayed, as expected, an approximate doubling of the original size, which also poses certain difficulties to the majority of classification algorithms. Hence, we experimented with an ensemble classification algorithm known as Gradient Boosting Machines (GBM), which is theoretically proven to be able to cope with a large number of features [9]. According to the referenced work, "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". Moreover, apart from the high-dimensional nature of our task at hand, a class imbalance issue appeared since the distribution of sentiment in Web 2.0 resources often portrays signs of imbalance. Imbalanced datasets correspond to domains where there are many more instances of some classes than others. For example, there are frequent occasions in political or religion discussions about a given topic over social media that generally present a skewed polarity distribution. In order to confront that phenomenon, a hybrid technique that performs a modification over a former technique [10] was applied. More specifically, we applied under-sampling to the instances that belong to the majority class and over-sampling to the minority class, respectively.

We experimented with numerous well-known algorithms using the initial Greek-only feature set (obtained upon preprocessing with basic filters, such as tokenization and stemming), as well as the enhanced translated one and also some basic feature reduction techniques, such as principal component analysis [11] and feature selection. Through extensive experimental evaluations, we found that GBM are superior to any other implementation. Additionally, when coping with the class imbalance issue, experimental results justify the use of the bootstrapping method since traditional algorithms favor the majority class in their predictions. GBM was again the best choice in the latter case, i.e., the imbalanced dataset.

This paper is organized as follows: Section 2 deals with related work on this domain, while Section 3 provides brief insight on GBM. Section 4 discusses the main rationale for using the translation of Greek terms as a feature generation technique. Section 5 describes the data formulation and manipulation steps, along with the experimental setup process, and finally, Section 6 presents the outcomes of empirical experimental evaluations. Section 7 concludes with some revision remarks about the contribution of this article and its main findings.

2. Related Work

Throughout recent years, a vast number of articles studying different types of sentiment analysis in English documents has been observed. Examples of such types include objectivity and subjectivity detection, opinion identification, polarity classification, entity or aspect-based sentiment classification, etc. Detailed insights into the aforementioned approaches can be found in survey papers authored by Pang and Lee [12], Liu and Zhang [13], as well as Mohammad [14]. However, only a few studies address the problem of sentiment analysis in Greek. In the following paragraphs, the related work

is subdivided into a brief outline on current methods for sentiment analysis with emphasis on social media content, followed by relevant research on Greek texts and, finally, discussing the use of resources from another language, i.e., multilingual sentiment analysis.

2.1. Sentiment Analysis and Social Media

When dealing with information services, social events and e-commerce, the human urge to populate thoughts has resulted in the great popularity of opinionated textual content. For example, taking the fact that the most common type of message on Twitter is about "me now" [15], it is evident that users frequently externalize their current state of mind, their emotions and their mood. Bollen et al. [16] argued that users express both their own mood in tweets about themselves and more generally in messages about other subjects. Additionally, many users, upon reading an article or buying a product, feel the need to share their opinion online about this [17]. We can safely say that a great part of information generated, but not limited, by specific resources is consumed and commented on positively or negatively in Web 2.0 content. In the same article (i.e., [17]), it is estimated that about 19% of microblog messages include a brand name, and 20% contains sentiment.

The research of [12] focused on more traditional ways of automatic sentiment analysis techniques. We can categorize sentiment analysis approaches into two classes: the techniques that are based on some polarity lexica (e.g., [18]) and the methods that are based on machine learning (e.g., [19]). In the former case, pre-compiled terms have been collected and annotated with an a priori sentiment score, which can be aggregated in order to extract the sentiment of a document or a sentence. Moghaddam and Popowich [20] constitute the polarity of product reviews by identifying the polarity of the adjectives that appear in them, with a reported accuracy of about 10% higher than traditional machine learning techniques. Nevertheless, such relatively high-precision techniques often fail to generalize when shifted to other, new domains or text formats, because they are not flexible regarding the ambiguity of sentiment terms. A significant number of lexicon-based methods has portrayed the benefits of contextual information [21,22] and has also indicated specific context words with a high impact on the polarity of ambiguous terms [23]. A frequent drawback of such solutions is the time-consuming and painstaking procedure of forming these polarity dictionaries of sentiment terms, despite the fact that there exist solutions in the form of distributed techniques.

The latter class of sentiment analysis methods, that of machine learning, operates by extracting syntactic and linguistic features [24,25] from text corpora that have been manually annotated as regards their sentiment. Subsequently, classification algorithms attempt to construct computational models of the separation boundary between the positive and negative sentiment. Certainly, there is research that performs a combination of these two trends (e.g., [25]). The machine learning techniques can be divided into supervised approaches like naive Bayes, decision trees, Support Vector Machines (SVM) ([26,27]), and unsupervised approaches [28], like pattern-logic classification according to a lexicon. It should be mentioned that in [29], a naive Bayes algorithm was found to perform better than many other typical classifiers.

Pak and Paroubek [24] aimed to classify random tweets by building a binary classifier, which used n-grams and POS features. Their model was trained on instances that had been annotated according to the existence of positive and negative emoticons (i.e., pictorial representations of a facial expression using punctuation marks, numbers and letters). Similar methods have been also suggested by [30], which also used unigrams, bigrams and POS features, though the former proved through experimental analysis that the distribution of certain POS tags varies between positive and negative posts. One of the reasons for the relative lack of linguistic techniques for opinion mining on social media is most likely due to the inherent difficulties in applying standard Natural Language Processing (NLP) techniques on low quality texts, something that machine learning techniques can, to some extent, overcome with sufficient training data. A characteristic example that demonstrates the above is that the Stanford Name Entity Recognizer (NER) drops from a 90.8% F-measure to merely 45.88% when applied to a set of tweets [31].

Recent advances in the domain introduce the use of semantics as an attempt to enhance existing features with semantic ones. By using linked data, ontologies and various lexical online resources, studies, such as [32–34], portray the advantages of having a semantic framework for identifying opinion holders and building sentiment scoring algorithms on its top. Note however that in order to include high-level semantics in sentiment analysis projects, lexical resources such as SentiWordNet, YAGO (Yet Another Great Ontology) ConceptNet, etc., are required, being for now mostly available in English.

2.2. Sentiment Analysis for Greek Texts

Greek is a language with limited, freely-available, linguistic resources. Therefore, most research on sentiment analysis for Greek relies on handcrafted sentiment lexica. For example, the authors of [35] construct an emotional sentiment lexicon for Greek and apply it on a Twitter dataset, created by [36]. The entries of this sentiment lexicon were collected through crawling the electronic version of the Greek dictionary by Triantafyllides [37]. The work of [38] also describes the use of a manual sentiment lexicon for Greek, containing about 27,000 types of positive words and 41,000 types of negative words. The authors also presented an open-source tool that inflects words in a semi-automatic manner. Upon creation of the lexicon, they used a simple bag-of-words approach that aggregated the sentiment score over all of the words within a document. A set of 1800 evaluations from the Greek version of TripAdvisor was used, and SVM was applied as the classification method.

A commercial software for polarity detection for entities and sentiment analysis, called "OpinionBuster", is presented in [39]. They presented a name entity recognition module that has been trained to locate entities from the reputation management domain, such as political parties, products of particular vendors and their competitors and perform sentiment analysis using Hidden Markov Models (HMM). Their corpus consisted of about 1300 RSS feeds from Greek political newspapers (Kathimerini and Real News). In [40], sentiment analysis was again performed on Twitter data, considering two milestones during the 2012 Greek elections, i.e., one week before and one week after. The goal of this work was to study the alignment between web and actual political sentiment in a bi-directional manner: the impact/reflection of the tweets' sentiment on the elections, as well as the impact of the elections' results on web sentiment and its shift around such a major political event. The authors examined the sentiment tagging in a supervised environment. Their hypothesis was focused on the positive vs. negative distinction, using statistical techniques such as count and frequency distributions. The alignment between actual political results and web sentiment in both directions was investigated and confirmed.

Finally, in [41], a framework for the lexicon-grammar of verb and noun predicates denoting emotion is presented, followed by its transformation into grammatical rules. The authors discuss the lack of significant resources and NLP preprocessing tools and, towards this direction, propose the enriching, re-purposing and re-using of already available Language Resources (LR) for the identification of emotion expressions in texts.

2.3. Multilingual Sentiment Analysis

Sentiment analysis research has predominantly been applied on English. Approaches to improve sentiment analysis in a resource-poor focus language in a multilingual manner include either: (a) the translation of resources, such as sentiment labeled corpora and sentiment lexicons, from English into the focus language, and use them as additional resources in the focus-language sentiment analysis system; or (b) the translation of the focus language text into a resource-rich language, such as English, and apply a powerful sentiment analysis system on the translation.

Initial study on multilingual sentiment analysis has primarily addressed mapping sentiment resources from English into morphologically complex languages, i.e., the former direction (a). Mihalcea et al. [42] used English resources to automatically generate a Romanian subjectivity lexicon using an English-Romanian dictionary. The work of [43] investigated both (a) and (b) for Arabic. More

specifically, they conducted several experiments by either using translated English polarity lexica into Arabic texts or freely-available machine translation engines and manual translations for converting Arabic to English in order to subsequently apply sentiment analysis.

The work of Politopoulou and Maragoudakis [27] was the first to introduce the idea of machine-translated aid towards sentiment analysis of Greek texts. In their approach, they translated the whole document, which led to significant deterioration of the original meaning, probably due to the noisy content (the instances were collected by a social media platform) and the early versions of the Greek-to-English machine translation engines. Balahur and Turchi [44] investigated the performance of statistical sentiment analysis on machine-translated texts. Opinion-bearing English phrases from the New York Times dataset were fed into an English sentiment analysis system that portrayed a prediction accuracy of approximately 68%. Following, the dataset was automatically translated into German, Spanish and French using publicly available machine-translation APIs, such as Google, Bing and Moses. The translated test sets were subsequently manually corrected for errors. Upon corrections for German, Spanish and French, a sentiment analysis system was trained on the translated training set for that language and tested on the translated-and-corrected test set. The authors observed that these German, Spanish and French sentiment analysis systems performed similar to the initial English sentiment classifier. There also exists research on using sentiment analysis to improve machine translation, such as the work by Chen and Zhu [45], but that is beyond the scope of the proposed work.

3. Gradient Boosting Machines

3.1. Boosting Methods Overview

The main characteristic of "boosting" is the conversion of a set of weak learners to a strong and robust classifier. A weak learner is practically any prediction model with a relatively poor performance (e.g., in terms of accuracy) that leads to unsafe conclusions, thus making it unusable due to the high rate of misclassification error. To convert a weak learner to a strong one, the predictions of a number of independent weak learners have to be combined. This combination is accomplished by taking the majority vote of every prediction of all weak learners as the final prediction. Another way to produce a strong learner from weak ones is to use the weighted average. A weak learner is added iteratively to the ensemble until the ensemble provides the correct classification. The most common representatives of the boosting family are Adaptive Boosting (AdaBoost), Gradient Boosting (GBM) and XGBoost (eXtreme Gradient Boost). These types serve as the base for a large number of boosting variations.

3.2. How the Gradient Boosting Algorithm Works

The common boosting techniques, like AdaBoost, rely on simple averaging of models in the ensemble. The main idea of boosting is to add new models to the ensemble sequentially. At each iteration, a new weak, base learner model is trained with respect to the error of the whole ensemble learnt so far. In the case of gradient boosting, the learning method successively fits new models to deliver a more accurate estimate of the class variable. Every new model is correlated with the negative gradient of the loss function of the system and tends to minimize it. This materializes by using the gradient descent method.

3.3. Gradient Descent Method

The gradient descent is an optimization algorithm that finds the local minimum of a function using steps proportional to the negative gradient of the function. Suppose there is a multi-variable function $F(x)$, defined and differentiable in a neighborhood of point a. $F(x)$ decreases fastest if one goes from a in the direction of the negative gradient of F at a, i.e., $\nabla F(a)$. For any point b where $b = a - \gamma \nabla F(a)$, if γ is small enough, it turns out that $F(a) \geq F(b)$. Therefore, if we consider a random starting point x_0 for a local minimum of F and apply the previous formula to a sequence of points, $x_0, x_1, \ldots, x_n$, such that $x_{n+1} = x_n - \gamma_n \nabla F(x_n)$ n ≥ 0, we obtain $F(x_0) \geq F(x_1) \geq F(x_1) \geq \ldots$, and

the sequence converges to the desired local minimum. This method is applicable to both linear and non-linear systems.

3.4. Application of the Gradient Descent Process to the Error Function

Before applying the gradient descent process to an error function, one should express the estimating function of a system F for an expected loss function g. For a given dataset in the form of X_i, Y_i (i = 1, ... n), X_i is a variable with k-dimensions and Y_i is a response with continuous or discrete values (in the case of continuous values, the problem is characterized as regression, in the case of discrete values as classification). Let us assume for reasons of simplicity that Y is univariate. From the function: $F: \mathbb{R}\kappa \rightarrow \mathbb{R}$ and expected loss g, the aim is to minimize the expected loss $\mathbb{E}[g(Y,F(X))]$, $g(.,.) : \mathbb{R}\, X\, \mathbb{R}+$.

$$\hat{F}(x) = arg \underset{F(x)}{min}\ g(y, F(x))$$

As mentioned earlier, F and g are the estimation and the loss function, respectively. The function F expresses the correlation of X, Y. To make sure that the method will work correctly, the loss function g should be smooth and convex in the second argument. The loss function depends on the domain with which we have to cope. The features of the system (e.g., if the system has outliers or it is high dimensional) pose a significant impact on the decision of how the machine learning task should be considered. In most of the cases, the loss function belongs to one of the following [9]:

(a) Gaussian L2
(b) Laplace L1
(c) Huber with δ specified
(d) Quantile with α specified
(e) Binominal
(f) AdaBoost
(g) Loss function for survival models
(h) Loss function counts data
(i) Custom loss function

In case the response of the system is continuous, we try (a) to (d) from the above list; in case the response is categorical, we try € or (f). To the extent of our knowledge, the three remaining loss functions are currently not supported by any open-source implementation of GBM.

The following figure (Figure 1) illustrates two continuous loss functions:

(a) L2 squared loss function with generic format: $\Psi(y,f)_{L_2} = \frac{1}{2}(y-f)^2$ and
(b) Huber loss function with generic format:

$$\Psi(y,f)_{Huber,\delta} = \begin{cases} \frac{1}{2}(y-f)^2, & |y-f| \leq \delta \\ \frac{1}{\delta}\left(|y-f| - \frac{\delta}{2}\right), & |y-f| > \delta \end{cases} \quad (1)$$

for the case (I) δ = 1, (II) δ = 0.5 and (III) δ = 0.25.

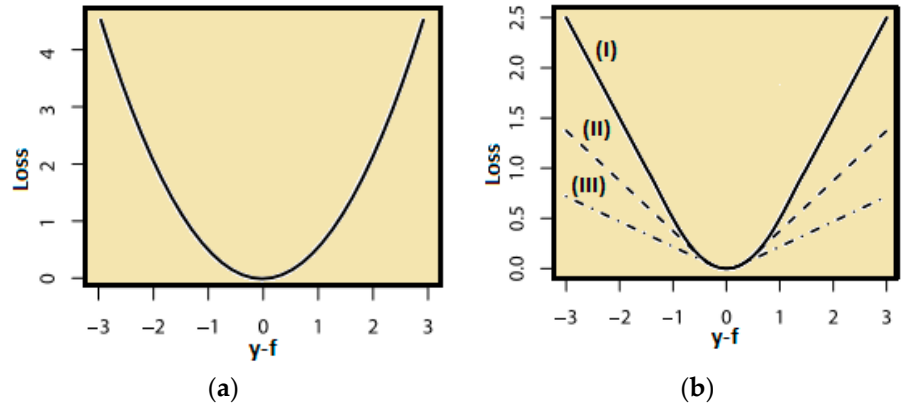

Figure 1. (**a**) Continuous loss function L2 squared; (**b**) Huber loss function.

3.5. Regularization for Base Learners

One of the most powerful characteristics that GBM has is the ability to use different types of base learners, like linear models, smooth models, decision trees, Markov models, etc. A list that categorizes the base learners is presented below:

1. Linear models: ordinary linear regression, ridge penalized linear regression and random effects.
2. Smooth models: P-splines, radial basis functions.
3. Decision trees: decision tree stumps, decision trees with arbitrary interaction depth.
4. Other cases of models: Markov random fields, custom base learner functions, wavelets.

This feature provides the ability of one GBM to encompass more than one base learners. In practice, this means that at the same time for the system we work on, we can have one function that includes a combination of base learners, e.g., decision trees and linear models. Therefore, it is feasible that complex models are created. At this point, we should mention that regardless of what base learners will be used in order to produce a satisfactory model, the regularization part is critical. There are explicit techniques that can be applied, which depend on the particular case. These techniques include, but are not limited to, subsampling, shrinkage and early stopping. Especially in early stopping, it prevents the model from overfitting. The fine-tuning process should definitely be a part of the model creation process. Additionally, GBM can cope with high-dimensional data because of the ability to create sparse models. This advantage is very useful in the case of sentiment analysis. Nevertheless, the flexibility of the algorithm has some objective disadvantages. GBM creates models by doing a great number of iterations, thus requiring plentiful resources in processing power and memory consumption, because every model is stored in memory.

3.6. Variable Importance

Previously, we had categorized the GBM algorithm as an ensemble algorithm. Like every ensemble algorithm, GBM is also capable of calculating the importance of each variable. An internal mechanism provides the rate at every single feature. From this rate is produced a list with all of the features with their relative importance. The order of variable importance can be used as a tool for exploratory data analysis and domain knowledge extraction, especially in cases for which the system is very complex.

4. Translation as a Feature Generation Method

In text mining, the traditional Bag Of Words (BOW) approach is inherently restricted to model pieces of textual information that are explicitly mentioned in the documents, provided that the same vocabulary is consistently used. Specifically, this approach has no access to domain knowledge possessed by humans and is not capable of dealing with facts and terms not mentioned in the

training set in a robust manner. We will borrow an example mentioned in [46], in order to exemplify the disadvantages of BOW. Consider Document #15264 in Reuters-21578, which is one of the most commonly-used datasets in research about text categorization. This document describes a joint mining venture by a consortium of companies, belonging to the category "copper". By examining the document, one can clearly observe that it mainly mentions the mutual share holdings of the companies involved (Teck Corporation, Cominco and Lornex Mining) and only briefly reports that the aim of the venture is mining copper. As a result, three popular classifiers, such as SVM, k-NN and decision trees, failed to classify the document correctly.

A possible solution to the aforementioned problem could involve the consideration of additional features, with the anticipation that these will be more informative than the existing ones. This strategy is not uncommon in machine learning and is referred as feature generation [47]. The motivation in feature generation is to search for or produce new features that describe the target concept better than the ones supplied with the training instances. Works by [48,49] described feature generation algorithms that led to substantial improvements in classification. In text mining, feature generation has been applied in terms of finding concepts of words either from WordNet [50] or from Open Directory and Wikipedia [51].

Our approach is also based on feature generation, but instead of finding synonyms or hypernyms through WordNet, we apply machine translation over Greek tokens. Certainly, this approach augments the total number of features, which will likely intensify the "curse of dimensionality" issue. Nevertheless, there exist feature selection algorithms, either inherent in classifiers, such as GBM or SVM, or individual, such as Pearson correlation. As already mentioned, we apply GBM, which can measure the influence of features by measuring the effect of each split on a variable in a decision tree to the log likelihood of the whole ensemble across all trees.

A research question emerged from our decision to utilize the translation of the Greek texts into English, i.e., whether to translate at the sentence or at the word level. By experimenting with both approaches over a small sample of the dataset and also by inspecting various instances, we observed that the noisy nature of such documents set insurmountable obstacles to the translation, resulting in very poor outcomes. An illustrative example, demonstrating the fact that using the Google translation API at the sentence level does not always give better results than translating at the word level, especially for content that is populated in Web 2.0 platforms, such as in our case, is provided below. Table 1 presents an initial negative comment about the former prime minister of Greece, as posted on the web. Furthermore, we provide its meaning in English and, finally, the outcome of the translation engine. One can clearly evaluate the translation performance at the sentence level over the general meaning; therefore, it is evident that in noisy environments, such as social media content, considering the full translation would not significantly aid the feature generation process compared to taking each token separately into account.

Table 1. Dealing with machine translation at the sentence level on a real example from our database.

Original Greek Text	Meaning (in English)	English Translation as Extracted from Google Translate®
στην περιπτωση του σαμαρα, ισχυει η ρηση, το μη χειρον, βελτιστον!..ομως δεν ηταν ουτε αυτοσ αξιοσ μιασ ελλαδασ...μονο και μονο απο το υφακι του εχασε, το στυλακι, εγω, εγω, εγω εκανα, αμετρητεσ φορεσ ελεγε εγω, εγωπαθησ και εγωισταροσ ..συν το οτι κυνηγησε τοσο αντισυνταγματικα ενα κομμα ολοκληρο, κατα τη γνωμη μου πολυσ κοσμοσ ειδικα δεξιοι τον μισησαν απο αυτο	In the case of Samaras (note: former PM of Greece) the famous saying "choose the lesser of two evils" applies! But he was also unworthy of Greece, he lost popularity only by his attitude and his style, countless times he was saying "I", he is an egomaniac and a great selfish person, plus that he prosecuted in an un-constitutional manner a political party, according to my opinion, a lot of people, particularly those that belong to the right political party hated him for this.	In the case of samarium apply the dictum, the non-chiral, Best! .. But was not he nor worthy of a ... Of Greece just from the yfaki missed the stylaki, he, I, I I did, it said countless times ego, egomaniac and egoistaros ..syn in that hunt both unconstitutional a party whole, kata my opinion, a lot of people especially hated him right from this

5. Experimental Setup

5.1. Data Collection

In the present study, the corpus was retrieved from user opinions posted in online Greek newspaper articles. These articles were posted at the online portal of the "Proto Thema" newspaper, which is the top selling Sunday paper in Greece, exceeding 100,000 prints each Sunday. The thematic areas of the articles varied from politics, society, finance and sports. Seven hundred forty instances were collected and annotated by two individuals, reaching a 98% inter-annotation agreement level. The only restriction was that each selected comment could be clearly classified as negative or positive. The initial format of the corpus was about 35% to 65% with regards to the analogy of positive and negative. It was very difficult to retain the balance since there were topics that most comments were biased mostly to the negative class (e.g., in Greek articles about migration and the economic crisis). However, our aim was to study the behavior of GBM to sentiment analysis in normal circumstances and in extreme ones, such as the imbalanced sets. Therefore, apart from the original corpus A, four additional datasets were derived from it, namely:

B. A mixed dataset, which consisted of the original Greek comments plus the English token augmentation.
C. A translated dataset, i.e., the original set upon translation in English and removal of Greek tokens. The reason for incorporating this set is to prove that the main approach of adding English translations performs better than the traditional approach of just translating the text into English.
D. An imbalanced to "positives" dataset, derived from the original Greek comments plus the English translation, but with removing most of the "negative" instances. The proportion was kept at about 95% positives and 5% negatives.
E. An imbalanced to "negatives" dataset, which consisted of the original Greek comments plus the English translation, having removed most of the "positive" cases. The proportion was also about 95% negatives and 5% positives.

5.2. Data Preprocessing

In text mining, the initial step is usually the transformation of free text into a structured vector representation, able to be analyzed by most of the machine learning algorithms. Since the collected data were not in the vector form we need them to be, a certain number of preprocessing steps needed to be carried out. The articles contained URL addresses that had been removed using an HTML tag identifier. The following steps describe the whole process.

(a) Tokenization: All stop-words have been removed, and all letters have been lowercased.
(b) Stemming: for the Greek set (denoted as A), an implementation of the [52] stemmer was applied.
(c) Translation: Each Greek token was translated by using the Google® Translator API.
(d) Creation of the document-term matrix: The number of rows of the matrix was equal to the number of the instances, and the number of columns was equal to the distinct number of Greek tokens plus the number of English tokens, as translated in Step c including the translation in the case of balanced data, more specifically 740 articles with translation, 260 positive and 480 negative. As previously explained, for the cases of Datasets D and E, i.e., the "positive" imbalanced and "negative" imbalanced sets, 5% of each class was kept within the document collection. On each cell of the document-term matrix, the value of the *tf-idf* weight of each single term was provided from 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
- $N(term)$: is a function that returns the number of documents in which the term appears.

(e) Reduce the dimensionality via Principal Component Analysis (PCA). PCA is a technique that reduces dimensionality by performing the transformation of possibly correlated variables to a fully new dataset with linearly uncorrelated variables, called principal components. Each principal component encodes variance starting from the first one, which is guaranteed to have the largest variance, meaning that it holds the biggest impact within data, compared to the other principal components, which also contain variance information in descending order of appearance. All of the principal components are actually the eigenvectors of the covariance matrix, which means that they are orthogonal. This was an optional step that aimed at alleviating the data sparsity problem and assisting traditional classifiers, such as SVM, decision trees, and naive Bayes. This step was not applied in the case of experimenting with GBM, since it is mathematically proven that GBM can cope with the large number of attributes.

(f) In order to assist classifiers, such as decision trees and naive Bayes, which are significantly influenced by high-dimensional vectors, another Feature Selection (FS) technique was applied apart from PCA, i.e., weighting each feature using the information gain criterion [53] and retaining the top-k of them. In the Experimental Results section, we describe the process of finding the optimal parameters for our classifiers, as well as the optimal number of principal components and the number k for the feature selection step.

The dimensionality (in terms of columns, i.e., features) for Datasets A and C, the original Greek set of opinions and their English translations upon the application of tokenization and stemming reached 7765 (it was about 11% more when stemming was not applied), while the number of features for Datasets B, D and E was 13,671.

5.3. Dealing with the Class Imbalance Problem

A common phenomenon in real-world applications, such as the one at hand, class imbalance problems often deteriorate the performance of traditional classifiers. By studying the current state of the art in the field, we have followed a hybrid approach that performs oversampling of the minority class based on the SMOTE (Synthetic Minority Oversampling TEchnique) [10] algorithm and under-sampling of the majority class using the Tomek links score to identify noisy and borderline examples. SMOTE is based on the idea that each example of the minority class is synthetically re-generated using its k-NN graph instead of a randomized sampling with replacement. The Tomek link [51] method can be viewed as the means for guided under-sampling, where the observations from the majority class are removed.

5.4. Evaluation Criteria and Performance

The performance of each method has been measured by using accuracy, precision and recall. Accuracy: $\frac{T_p+T_n}{T_p+T_n+F_p+F_n}$; precision: $\frac{T_p}{T_p+F_p}$; recall: $\frac{T_p}{T_p+F_n}$; where T_p, T_n. are true positive and true negative; these are the correct positive and correct negative predictions, respectively. F_p, F_n are false positive and false negative and are the wrongly positive and wrongly negative predictions, respectively. A 10-fold cross-validation approach has been used in order to evaluate the performance for every classifier.

The above flowchart (Figure 2) depicts the formation process of each dataset and the experimental evaluation procedure. As shown in the image, the original corpus is treated at the beginning with some basic preprocessing steps, while the idea of applying the English translation to the original set is depicted in the second row of the chart. Based on this augmentation, the mixed dataset is created. Furthermore, since we need to evaluate the proposed methodology in real-world scenarios, where it

is very common that one class dominates the other in distribution ratios, we have proceeded to the generation of two, highly imbalanced sets per each class label. Finally, for all described datasets, the PCA dimensionality reduction and the FS technique were optionally applied, to assist the traditional classifiers in recovering from data sparsity.

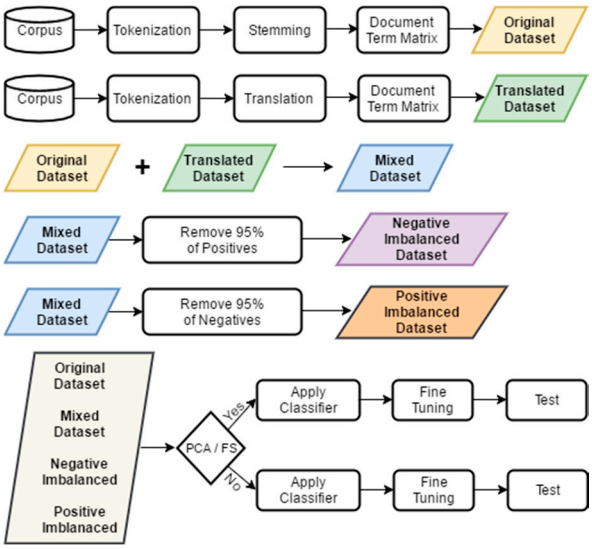

Figure 2. Methodology flowchart. FS, Feature Selection.

5.5. Classification Benchmark

In order to investigate the performance of the proposed GBM method and evaluate it against other, well-known classifiers that have been previously applied in sentiment analysis tasks with success, as explained in the Related Work section, the following classifiers have been incorporated: decision trees, naive Bayes, support vector machines and deep neural network learning (using recurrent neural networks). In the case of decision trees, naive 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 or the feature selection stage.

6. Experimental Results

The datasets we used as shown in Figure 2 were:

(a) The original dataset, which contains only the Greek tokens.
(b) The mixed dataset, which contains Greek tokens and their translation.
(c) The translated dataset (no Greek tokens).
(d) The negative imbalanced dataset, which was derived from the mixed dataset removing 95% of the positives.
(e) The positive imbalanced dataset, which was derived from the mixed dataset removing 95% of the negatives.

The underlying idea in these datasets is to approximate realistic conditions in opinion mining and sentiment analysis systems. The imbalanced datasets have been dealt with using the under-sampling and over-sampling technique mentioned above. Each of the aforementioned datasets also underwent the PCA data dimensionality technique and the feature weighting step using information gain. PCA and FS were applied to all used classifiers (decision trees, naive Bayes and support vector machines), except GBM.

The experiments were carried out on a rack server, with 2 Intel Xeon E5-2650/2 GHz 8-core processors, with 64 GB of RAM, running under Linux (CentOS). The RapidMiner® framework was

used throughout the experimental evaluations. In the cases of GBM and deep learning, the H_2O Open Source Fast Scalable Machine Learning API was called within RapidMiner. Since these algorithms are highly dependent on their parameters, optimization of GBM, deep learning, SVM, as well as the optimal number of principal components and, finally, the top-k features in the FS step was carried out using a grid search over numerous different parameter settings, upon evaluation on a held-out set and having the harmonic mean of precision and recall, named the F-measure, as the performance criterion during the optimization search. In order to reduce the training time, for the cases of PCA and FS, we experimented with five different parameter settings. We asked the PCA algorithm to retain five thresholds of variance, starting from 20% and climaxing at 60% using 10% incremental steps. Furthermore, if the FS weighting process, we set five different top-k values of features to be retained, i.e., 100, 250, 500, 750 and 1000. For the technical details of the optimization operators, please refer to the RapidMiner documentation (http://docs.rapidminer.com/studio/operators/modeling/optimization/parameters/optimize_parameters_grid.html), namely the "Optimize Parameters/Grid" operator. Since the following paragraphs describe experiments on a different dataset each time, the optimal parameters found at each set will be tabulated.

6.1. Experiment A

The initial experiment considered the original dataset with GBM against the other algorithms used in the benchmark, with and without the PCA and FS reduction. Table 2 tabulates the best outcomes for Dataset A, i.e., the original set of the Greek tokens only. These scores correspond to the PCA transformation, keeping 30% of the original variance, i.e., approximately 250 principal components. The best numbers per performance metric are shown in boldface.

Table 2. Performance of GBM and other benchmarking methods for Dataset A: original. GBM, Gradient Boosting Machine.

Dataset A: Original	Accuracy	Precision (+)	Recall (+)	Precision (−)	Recall (−)
Decision Trees	68.42%	59.12%	62.86%	75.76%	79.54%
SVM	77.82%	80.10%	67.60%	79.20%	87.40%
Naive Bayes	69.12%	58.40%	64.30%	76.30%	72.50%
Deep Learning	72.00%	67.50%	74.80%	82.40%	78.00%
GBM	**84.20%**	**82.40%**	**77.90%**	**86.30%**	**88.30%**

As observed, the original set displays quite satisfactory results using almost any algorithm. GBM is the most powerful method, outperforming all other methods in precision and recall of both classes, followed by SVM. GBM was found to outperform all others in all metrics by a significant difference that reached almost 12% when compared to deep learning. For reasons of space, we have not included the correspondent table when utilizing feature selection; however, we must note that FS was quite close to PCA and, thus, not proficient at significantly improving the performance of the classifiers. When incorporating FS, the results were slightly worse than the ones found in Table 2 by a varying factor of 1.5% to 6%, for various top-k settings. Table 3 presents the optimal parameters of the classification algorithms for the aforementioned results of Table 2.

Table 3. Parameter values, upon the completion of the parameter tuning process for Experiment A.

Algorithm	Parameters
Decision Trees	Criterion: Gini Index; Maximal Depth: 60; Confidence: 0.3; Minimal Leaf Size: 12
SVM	Kernel Type: polynomial; Degree: 3; C: 1.35; Epsilon: 10^{-4}
Naive Bayes	*No parameters to optimize*
Deep Learning	Activation: Tanh; Hidden Layers: 3; Loss Function: Quadratic; Epochs: 30;
GBM	Loss Function: Ada-boost; Number of Trees: 120; Max-depth: 4; Learning rate: 0.05

6.2. Experiment B

The second experiment dealt with the incorporation of Dataset B, i.e., the mixed dataset, using the translation technique. Table 4 tabulates the performance metrics per each algorithm. These figures were found to be the greatest when utilizing PCA transformation, retaining 30% of the variance of the mixed set.

Table 4. Performance of GBM against other benchmarking methods for Dataset B: Mixed.

Dataset B: Mixed	Accuracy	Precision (+)	Recall (+)	Precision (−)	Recall (−)
Decision Trees	70.50%	63.23%	66.60%	77.15%	86.50%
SVM	81.35%	83.50%	68.50%	83.00%	91.80%
Naive Bayes	71.80%	62.35%	67.27%	79.34%	75.06%
Deep Learning	75.70%	67.90%	81.30%	86.60%	78.50%
GBM	**87.85%**	**84.66%**	**83.30%**	**91.30%**	**91.75%**

An examination of the fourth table reveals that the methodology of including additional features improves the classification outcome for every algorithm. However, the improvement is clearly bigger when one uses methods for performing some sort of feature selection, such as GBM and deep learning. For the former, we also observe that it has reached almost 88% in accuracy, and simultaneously, it retains high precision and recall values for both classes, while in other methods, usually the recall of the positive or negative class often varies from the other metrics. For example, in deep learning, the positive precision is 67.9%, while the recall reaches 81.3%. In other words, the algorithm attempted to fit almost all positive examples, but in this process, it was not able to generalize, thus labeling some negative examples as positives. Results were slightly worse when utilizing the FS process, with the case of keeping the top-750 features being approximately 2% to 3% lower than PCA.

Figure 3 contains a graph depicting the improvement of the inclusion of the proposed technique with the translation of Greek tokens. As explained before, GBM and deep learning present the best improvement, reaching almost 7% better outcomes than when considering the original dataset. An interesting observation is that the mixed dataset aids all other methodologies. This might sound peculiar since doubling the initial feature set means additional load for any classifier; however, note that the translation is not a simple doubling of the size of vectors, since it encodes semantic relations between each token and the corresponding class distribution, making the decision boundaries easier to recognize.

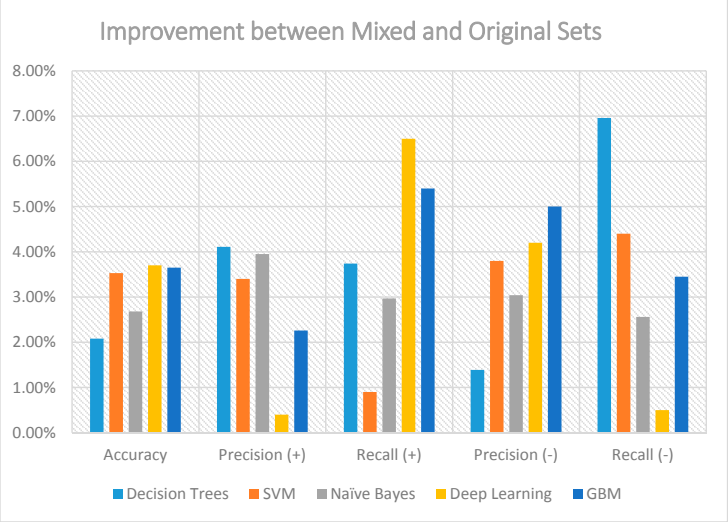

Figure 3. Improvement when considering the mixed dataset, in terms of accuracy, precision and recall per each class label, over the original dataset.

In analogy with the previous experimental scenario, we tabulate the optimal hyper-parameters for the involved algorithms in the following Table 5.

Table 5. Parameter values, upon completion of the parameter tuning process for Experiment B.

Algorithm	Parameters
Decision Trees	Criterion: Gini Index; Maximal Depth: 40; Confidence: 0.25; Minimal Leaf Size: 20
SVM	Kernel Type: polynomial; Degree: 2; C: 3.9; Epsilon: 2×10^{-4}
Naive Bayes	*No parameters to optimize*
Deep Learning	Activation: Rectifier; Hidden Layers: 2; Loss Function: Quadratic; Epochs: 20;
GBM	Loss Function: Ada-boost; Number of Trees: 100; Max-depth: 6; Learning rate: 0.35

6.3. Experiment C

The third experiment dealt with the inclusion of Dataset C, i.e., the translated dataset, without any Greek tokens. Table 6 tabulates the performance outcomes per classifier. The best setting was found to be achieved when considering the PCA transformation, retaining 40% of the variance of the initial Dataset C.

Table 6. Performance of GBM against other benchmarking methods for Dataset B: mixed.

Dataset B: Mixed	Accuracy	Precision (+)	Recall (+)	Precision (−)	Recall (−)
Decision Trees	69.20%	59.68%	62.90%	75.88%	78.22%
SVM	76.58%	79.40%	68.30%	78.74%	86.90%
Naive Bayes	69.65%	58.58%	65.25%	76.75%	73.00%
Deep Learning	72.10%	68.12%	75.30%	82.60%	77.90%
GBM	**83.90%**	**82.10%**	**78.15%**	**86.47%**	**88.20%**

A closer look at the results reveals that the translated dataset is slightly better than the original one (i.e., the Greek texts only), but with a minor improvement of about 0.5% to 0.9%. Furthermore, a comparison of the table above with the outcomes of the mixed Dataset B shows that the proposed methodology of augmenting the initial set with the English translations is the most beneficial in terms of precision and recall metrics. Therefore, one could claim that a simple translation of Greek into English without considering both languages results in lower classification performance. Additionally, even in this case, GBM performs noticeably better that any other classifier. Results were slightly worse when utilizing the FS process, with the case of keeping the top-500 features being approximately 1.5% to 2.7% lower than PCA.

As with the previous cases, we tabulate the optimal hyper-parameters for the involved algorithms in the following Table 7.

Table 7. Parameter values, upon completion of the parameter tuning process for Experiment C.

Algorithm	Parameters
Decision Trees	Criterion: Gini Index; Maximal Depth: 50; Confidence: 0.3; Minimal Leaf Size: 30
SVM	Kernel Type: polynomial; Degree: 2; C: 2.4; Epsilon: 2×10^{-4}
Naive Bayes	*No parameters to optimize*
Deep Learning	Activation: Tanh; Hidden Layers: 2; Loss Function: Quadratic; Epochs: 40;
GBM	Loss Function: Ada-boost; Number of Trees: 120; Max-depth: 5; Learning rate: 0.3

6.4. Experiment D

The fourth experiment was about assessing the influence of class imbalance on GBM, as well as on the other set of classifiers. For that reason, we considered both imbalanced sets, namely the C positive imbalanced and the D negative imbalanced set. In order to demonstrate that the proposed strategy

of oversampling the minority class and under-sampling the majority is actually beneficial, we first tabulate the performance metric without using it and subsequently when including it. Table 8 arranges the outcomes of the experimental run on Dataset C positive imbalanced, without any handling of the class imbalance issue. Again, these outcomes were obtained when utilizing the PCA method, for 20% of the initial variance.

Table 8. Performance of GBM against other benchmarking methods for Dataset C positive imbalanced without any measures for the class imbalance problem.

Dataset C: Imbalanced +	Accuracy	Precision (+)	Recall (+)	Precision (−)	Recall (−)
Decision Trees	96.21%	2.50%	16.97%	98.02%	98.22%
SVM	61.5%	5.2%	74.1%	98.3%	63.8%
Naive Bayes	77.75%	9.1%	65.7%	98.8%	74.3%
Deep Learning	75.15%	8.50%	69.20%	99.5%	76.53%
GBM	**88.20%**	**21.5%**	**75.1%**	**99.6%**	**97.9%**

An initial observation is that in this case, accuracy is actually not informative since it may appear high, but notice that the minority class label displays very poor results, meaning that almost all positive instances were incorrectly classified. Even though GBM again outperforms all other approaches, the 20% precision in the positive class is certainly not desirable. Thus, the following Figure 4 illustrates the results when considering the measures explained earlier, for dealing with class imbalance.

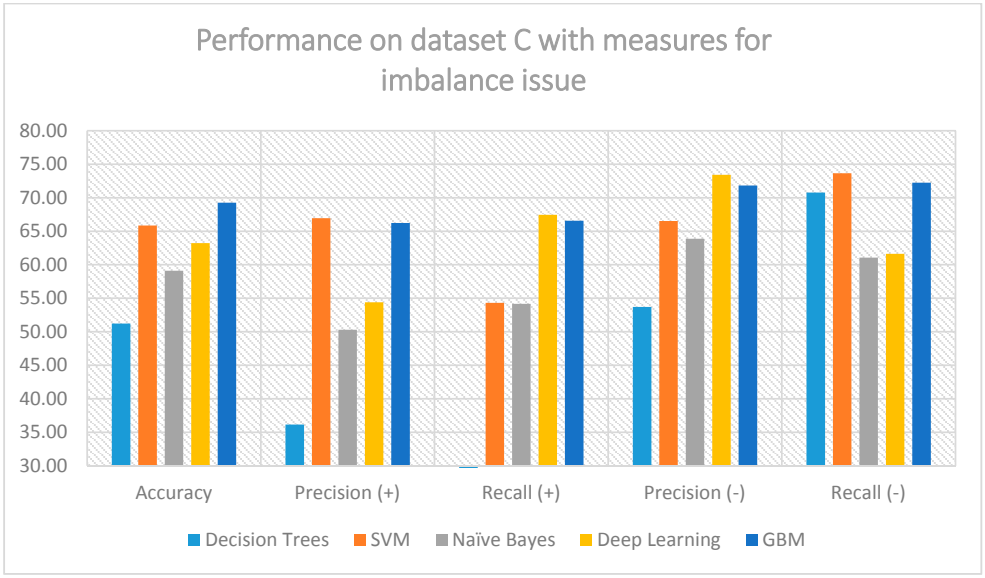

Figure 4. Evaluation scores of all classifiers against the positive imbalanced set upon taking measures for the class imbalance problem.

The improvement of the positive class precision and recall is evident. From the initial range of 2% to 20% for the precision metric, upon applying the measures we discussed earlier, we can observe that GBM has reached almost 67.5% precision and 65% recall. The same behavior is also apparent for other classifiers, meaning that the measures paid off, since the minority class is not confronted with success.

Table 9 tabulates the outcomes of the experimental run on Dataset D negative imbalanced, without any handling of the class imbalance issue.

As seen from the above table, the minority class (which is now the negative sentiment) also suffers from deteriorated metrics, both in precision and recall. It is therefore evident that in real-world sentiment analysis tasks, where class imbalance is a common phenomenon, even the most sophisticated

classifier and even considering advanced feature selection techniques cannot guarantee the creation of a robust prediction model. As with the previous dataset, Figure 5 represents the outcome of the evaluation, this time by considering the sub-sampling measures for class imbalance.

Table 9. Performance of GBM against other benchmarking methods for Dataset D negative imbalanced without any measures for the class imbalance problem.

Dataset D: Imbalanced −	Accuracy	Precision (+)	Recall (+)	Precision (−)	Recall (−)
Decision Trees	77.23%	95.20%	79.9%	16.9%	43.30%
SVM	62.25%	95.94%	56.43%	11.8%	65.20%
Naive Bayes	56.33%	90.79%	58.13%	6.46%	33.10%
Deep Learning	71.80%	98.00%	73.45%	17.33%	79.50%
GBM	**72.15%**	**98.7%**	**74.15%**	**19.9%**	**83.00%**

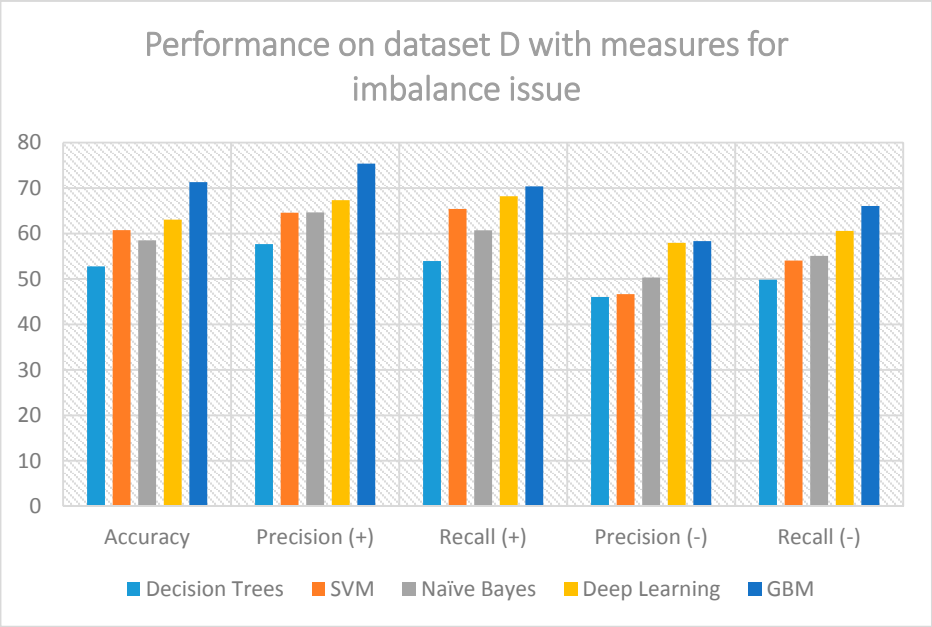

Figure 5. Evaluation scores of all classifiers against the negative imbalanced set upon taking measures for the class imbalance problem.

The improvement in the negative class (both precision and recall) is apparent when compared to the previous table. Notice that precision was initially varying from 6.5% to 19.9%, and now, it has reached almost 58.5%. Finally, we should mention that GBM was constantly found to be the most robust classification method, in all forms of datasets, with and without measures for the class imbalance problem. The parameters for the third experiment, either for the positive or the negative imbalanced sets, are tabulated below, on Table 10.

Table 10. Parameter values, upon completion of the parameter tuning process for Experiment D.

Algorithm	Parameters
Decision Trees	Criterion: Gain Ratio; Maximal Depth: 30; Confidence: 0.2; Minimal Leaf Size: 4
SVM	Kernel Type: radial; kernel gamma: 1.8; C:0.5; Epsilon: 4×10^{-4}
Naive Bayes	*No parameters to optimize*
Deep Learning	Activation: Rectifier; Hidden Layers: 2; Loss Function: Huber; Epochs: 50
GBM	Loss Function: Ada-boost; Number of Trees: 60; Max-depth: 4; Learning rate: 0.25

7. Conclusions

The presented research, which is an extension of [54], studied the use of machine learning in sentiment analysis tasks, particularly in situations where the text is given in a language with a relatively minimal set of linguistic analysis gazetteers and modules, such as the POS tagger, syntactic shallow parsers, etc. Modern Greek has its place to this category of under-resourced languages, and the present work collected real-world sentiment data, obtained from Web 2.0 platforms, and followed an idea of using machine translation of Greek tokens as a feature generation step. More specifically, instead of utilizing a hand-crafted polarity lexicon for Greek, which would have insufficient impact on the accuracy of the predictions due to the noisy social media linguistic style, we proposed a method that considers the translation of each Greek token as an additional input feature. Even though this process may appear to bring additional effort and complexity to the majority of classification algorithms, the use of gradient boosting machines, a robust ensemble method that can handle sparsity in high-dimensional data, appeared to be valuable for the task at hand, outperforming a family of well-known methods for sentiment analysis. Even when confronted with other state of the art classifiers, such as deep neural networks (from the family of recurrent neural networks), the proposed method demonstrated superior performance. Moreover, since sentiment data in real cases present a high level of class imbalance between the positive and the negative label, we applied sophisticated sampling methods for not allowing the bias of the classifiers towards the majority class. Yet again, GBM was found to be the superior solution in terms of precision and recall per each class label. In the future, we are oriented towards using semantic features, such as topic models, to further improve the sentiment analysis models.

Author Contributions: M. Maragoudakis. conceived and designed the experiments; V. Athanasiou and M. Maragoudakis performed the experiments; M. Maragoudakis. analyzed the data; V. Athanasiou and M. Maragoudakis wrote the paper. Authorship must be limited to those who have contributed substantially to the work reported.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Taylor, E.M.; Rodriguez, C.; Velasquez, J.D.; Ghosh, G.; Banerjee, S. Web Opinion Mining and Sentimental Analysis. In *Techniques in Web Intelligence-2, SCI 452*; Velásquez, J.D., Palade, V., Jain, L.C., Eds.; Springer: New York, NY, USA, 2012; pp. 105–126.
2. Maynard, D.; Bontcheva, K.; Rout, D. Challenges in developing opinion mining tools for social media. In Proceedings of the Workshop at LREC 2012, Istambul, Turkey; 2012; pp. 15–22.
3. Ravikant, N.; Rifkin, A. Why Twitter is Massively Undervalued Compared to Facebook. *TechCrunch*. 2010. Available online: https://techcrunch.com/2010/10/16/why-twitter-is-massively-undervalued-compared-to-facebook/ (accessed on 2 March 2017).
4. Skeels, M.M.; Grudin, J. When social net-works cross boundaries: A case study of workplace use of Facebook and LinkedIn. In Proceedings of the ACM 2009 International Conference on Supporting Group Work, (GROUP '09), Sanibel Island, FL, USA, 10–13 May 2009; pp. 95–104.
5. Abel, F.; Gao, Q.; Houben, G.J.; Tao, K. Semantic Enrichment of Twitter Posts for User Profile Construction on the Social Web. *ESWC* **2011**, *6644*, 375–389.
6. Mendes, P.N.; Passant, A.; Kapanipathi, P.; Sheth, A.P. Linked open social signals. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, (WI-IAT '10), Washington, DC, USA, 31 August–31 September 2010; pp. 224–231.
7. Han, B.; Baldwin, T. Lexical normalisation of short text messages: Makn sens a #twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, (HLT '11), Portland, OR, USA, 19–24 June 2011; pp. 368–378.
8. Gouws, S.; Metzler, D.; Cai, C.; Hovy, E. Contextual bearing on linguistic variation in social media. In Proceedings of the Workshop on Languages in Social Media, (LSM '11), Portland, OR, USA, 23 June 2011; pp. 20–29.

9. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2011**, *7*, 21. [CrossRef] [PubMed]
10. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.
11. Shlens, J. A Tutorial on Principal Component Analysis, Derivation, Discussion and Singular Value Decomposition. 25 March 2003. Available online: https://www.semanticscholar.org/paper/A-TUTORIAL-ON-PRINCIPAL-COMPONENT-ANALYSIS-Shlens/a99e0f8f58af7a91e26c1eda54e0cca3e3e03df3 (accessed on 2 March 2017).
12. Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 1–135. [CrossRef]
13. Liu, B.; Zhang, L. A survey of opinion mining and sentiment analysis. In *Mining Text Data*; Springer: London, OH, USA, 2012; pp. 415–463.
14. Mohammad, S.M. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion Measurement*; Meiselman, H., Ed.; Elsevier: Amsterdam, The Netherlands, 2016.
15. 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, Savannah, GA, USA, 6–10 February 2010; ACM: New York, NY, USA, 2010; pp. 189–192.
16. Bollen, J.; Pepe, A.; Mao, H. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *Comput. Sci.* **2009**.
17. Jansen, B.J.; Zhang, M.; Sobel, K.; Chowdury, A. Twitter power: Tweets as electronic word of mouth. *JASIST* **2009**, *60*, 2169–2188. [CrossRef]
18. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]
19. Boiy, E.; Moens, M.-F. A machine learning approach to sentiment analysis in multilingual web texts. *Inf. Retr.* **2009**, *12*, 526–558. [CrossRef]
20. Moghaddam, S.; Popowich, F. Opinion polarity identification through adjectives. *CoRR* **2010**, arXiv:1011.4623.
21. Weichselbraun, A.; Gindl, S.; Scharl, A. A context-dependent supervised learning approach to sentiment detection in large textual databases. *J. Inf. Data Manag.* **2010**, *1*, 329–342.
22. Wilson, T.; Wiebe, J.; Hoffmann, P. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Comput. Linguist.* **2009**, *35*, 399–433. [CrossRef]
23. Gindl, S.; Weichselbraun, A.; Scharl, A. Cross-domain contextualization of sentiment lexicons. In Proceedings of the 19th European Conference on Artificial Intelligence (ECAI-2010), Lisbon, Portugal, 16–20 August 2010; pp. 771–776.
24. Pak, A.; Paroubek, P. Twitter Based System: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives. In Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, CA, USA, 15—16 July 2010; pp. 436–439.
25. Zhang, L.; Ghosh, R.; Dekhil, M.; Hsu, M.; Liu, B. Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis Technical Report HPL-2011-89, HP 21 June 2011. Available online: http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html (accessed on 12 February 2016).
26. Politopoulou, V.; Maragoudakis, M. On Mining Opinions from Social Media, Communications in Computer and Information Science, Engineering Applications of Neural Networks. In Proceedings of the Lazaros Iliadis, Harris Papadopoulos, Chrisina Jayne, Halkidiki, Greece, 13–16 September 2013; pp. 474–484.
27. Maynard, D.; Funk, A. Automatic detection of political opinions in tweets. In Proceedings of the 8th International Conference on the Semantic Web (ESWC 2011), Heraklion, Crete, Greece, 29–30 May 2011; pp. 88–99.
28. Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–127. [CrossRef]
29. Go, A.; Bhayani, R.; Huang, L. *Twitter Sentiment Classification Using Distant Supervision*; Technical Report CS224N Project Report; Stanford University: Stanford, CA, USA, 2009.
30. Liu, X.; Zhang, S.; Wei, F.; Zhou, M. Recognizing named entities in tweets. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT '11), Portland, Oregon, 19–24 June 2011; pp. 359–367.
31. Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 855–868. [CrossRef] [PubMed]

32. Recupero, D.R.; Presutti, V.; Consoli, S.; Gangemi, A.; Nuzzolese, A. Sentilo: Frame-based sentiment analysis. *Cognit. Comput.* **2014**, *7*, 211–225.
33. Gangemi, A.; Presutti, V.; Reforgiato Recupero, D. Frame-Based Detection of Opinion Holders and Topics: A Model and a Tool. *IEEE Comput. Intell. Mag.* **2014**, *9*, 20–30. [CrossRef]
34. Reforgiato Recupero, D.; Consoli, S.; Gangemi, A.; Nuzzolese, A.G.; Spampinato, D. A semantic web based core engine to efficiently perform sentiment analysis. In *ESWC Satellite Events*; Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A., Eds.; Springer: New York, NY, USA, 2014; Volume 8798, pp. 245–248.
35. Kalamatianos, G.; Malis, D.; Arampatzis, A. Sentiment analysis of greek tweets and hashtags using a sentiment lexicon. In Proceedings of the 19th Panhellenic Conference on Informatics, Athens, Greece, 1–3 October 2015; ACM: New York, NY, USA, 2015.
36. Burnside, G.; Papadopoulos, S.; Petkos, G. D2.3 Social Stream Mining Framework. Available online: http://www.socialsensor.eu/images/D2.3.pdf (accessed on 2 March 2017).
37. Triantafyllides, G. *Dictionary of Standard Modern Greek*; Institute for Modern Greek Studies of the Aristotle University of Thessaloniki: Thessaloniki, Greece, 1998.
38. Markopoulos, G.; Mikros, G.; Iliadi, A.; Liontos, M. Sentiment analysis of hotel reviews in Greek: A comparison of unigram features of cultural tourism in a digital era. In *Springer Proceedings in Business and Economics*; Springer: New York, NY, USA, 2015; pp. 373–383.
39. Petasis, G.; Spiliotopoulos, D.; Tsirakis, N.; Tsantilas, P. Sentiment analysis for reputation management: Mining the Greek web. In *Artificial Intelligence: Methods and Applications*; Likas, A., Blekas, K., Kalles, D., Eds.; Springer: Heidelberg, Germany, 2014; Volume 8445, pp. 327–340.
40. Kermanidis, K.L.; Maragoudakis, M. Political sentiment analysis of tweets before and after the Greek elections of May 2012. *Int. J. Soc. Netw. Min.* **2013**, *1*, 298–317. [CrossRef]
41. Giouli, V.; Fotopoulou, A. Linguistically motivated language resources for sentiment analysis. In Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, Dublin, Ireland, 24 August 2014; p. 39.
42. Mihalcea, R.; Banea, C.; Wiebe, J. Learning multilingual subjective language via cross-lingual projections. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech, 23–30 June 2007.
43. Mohammad, S.M.; Salameh, M.; Kiritchenko, S. How translation alters sentiment. *J. Arti. Intell. Res.* **2016**, *55*, 95–130.
44. Balahur, A.; Turchi, M. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Comput. Speech Lang.* **2014**, *28*, 56–75. [CrossRef]
45. Chen, B.; Zhu, X. Bilingual sentiment consistency for statistical machine translation. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014; pp. 607–615.
46. Gabrilovich, E.; Markovitch, S. Feature generation for text categorization using world knowledge. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, UK, 30 July–5 August 2005; pp. 1048–1053.
47. Markovitch, S.; Rosenstein, D. Feature Generation Using General Constructor Functions. *Mach. Learn.* **2002**, *49*, 59–98. [CrossRef]
48. Hu, Y.; Kibler, D. A Wrapper Approach for Constructive Induction. *AAAI-96*. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.9922 (accessed on 2 March 2017).
49. Murphy, P.; Pazzani, M. ID2-of-3: Constructive Induction of M-of-N Concepts for Discriminators in Decision Trees. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.144.6995 (accessed on 2 March 2017).
50. Urena-Lopez, L.; Buenaga, M.; Gomez, J. Integrating linguistic resources in TC through WSD. *Comput. Hum.* **2001**, *35*, 215. [CrossRef]
51. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling imbalanced datasets: A review. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *30*, 25–36.
52. Ntais, G. Development of a Stemmer for the Greek Language. Master's Thesis, Department of Computer and System Sciences, Royal Institute of Technology, Stockholm University, Malmö, Sweden, 2006.

53. Azhagusundari, B.; Thanamani, A.S. Feature Selection based on Information Gain. *Int. J. Innov. Technol. Explor. Eng. (IJITEE)* **2013**, *2*, 18–21.
54. Athanasiou, V.; Maragoudakis, M. Dealing with High Dimensional Sentiment Data Using Gradient Boosting Machines. In Proceedings of the 12th IFIP WG 12.5 International Conference and Workshops, (AIAI 2016), Thessaloniki, Greece, 16–18 September 2016.