Sentiment Analysis for the Greek Language

Conference Paper · November 2016

DOI: 10.1145/3003733.3003769

4 authors, including:

Nikolaos Spatiotis

Dr Michael Paraskevas
Technological Educational Institute of Western Greece

Isidoros Perikos
University of Patras

Sentiment Analysis for the Greek Language

Nikolaos Spatiotis[1], Iosif Mporas[2], Michael Paraskevas[1,3], Isidoros Perikos[3,4]

[1]Computer and Informatics Engineering Dept., Technological Educational Institute of Western Greece, Greece
[2]Schoolof Engineering and Technology, University of Hertfordshire, United Kingdom
[3]Computer Technology Institute and Press "Diophantus", Patras, Greece
[4]Computer Engineering and Informatics Department, University of Patras, Greece
{mparask, perikos}@cti.gr

ABSTRACT

In recent years, the rapid development of technology has invaded dynamically in all areas of human activity by facilitating communication, information and interaction between people worldwide through Internet use. Internet users are no longer passive recipients of information, but new information data creators expressing ideas, opinions, feelings or their views on a service-product. Taking into account the evolution and use of mobile devices and the proliferation of wireless networks, the timely and widespread use of social networks and services satisfying the above uses are understandable. In this paper, we present an approach to analyze textual data in Greek language and extract meaningful information regarding the writer's opinion. More specifically, we present a supervised approach which classifies user generated comments into the proper polarity category. An extensive experimental study was conducted in the context of users' attitudes and opinions on e-lectures that they attended. The results were very promising, indicating that the approach was accurate and able to correctly classify opinions into the proper category.

Keywords

Sentiment analysis, machine learning, Text Mining, Sentiment classification, feature vectors

1. INTRODUCTION

The Sentiment Analysis (also known as opinion mining) refers to a broad field of Natural Language Processing (NLP), Computational Linguistics and Mining Text. The main methodology characteristic of the analysis is the Sentiment Classification.

Sentiment Classification is characterized as the technique of identification and emotional grouping of proposals which relates to the Opinion Mining from a text and is divided into individual fields in emotional orientation and in emotional classification. Regarding the first, it recognizes and classifies emotionally a proposal on the polarity (positive, negative and neutral) and the degree of intensity (low, medium and high). The second category i.e. of the emotional classification is examined and is categorized an opinion as to the objectivity or subjectivity that it dispose. The use of this taxonomy has been very popular for the last ten years.

Pang & Lee [1], [6] have made multiple researches regarding to sentiment analysis, sentiment classification, sentiment mining and opinion extraction. In one of them they proposed a new learning-machine method to determine the sense of polarity. The document polarity classification, as it is called, is a major challenge in this area. Cambria et al. [3], [5] in their research presented concept-level sentiment analysis providing new approaches for opinion mining and sentiment analysis not only from a simple word-level analysis but also from unstructured textual data. Liu [2] focused on subjectivity sentiment classification, sentiment lexicon generation analysis of comparative opinions and opinion summarization.

Our research focuses on opinion analysis of Greek language by creating a computer model of opinion processing and subjectivity that governs a text. For this research we used data derived from questionnaires collected by the Greek School Network in expressing their views Learners -who attended a web service (teleconference) - answering four open-ended questions (free text).The paper describes a Greek sentiment analysis system based on the way of writing texts written in the Greek language in order to be generated features vectors which together with classification algorithms to give us the opportunity classification of Greek texts based on their personal opinion and the degree of satisfaction. We created an automatic system that classifies opinions polarity into five classes reporting sufficiently performance and accuracy percentages.

The rest of the work is organized as follows: Section 2 presents background topics and Section 3 outlines the experimental setup. Section 4 describes the experimental study conducted and the results collected. Finally, conclusions and directions for future work are presented in Section 5.

2. Background Topics

2.1 Description of approaches

The Sentiment and Opinion Analysis can be applied at various levels of analysis and classified according to the way we approach the text and the desired detail of the extracted information [7].

Document / Text - layer: At this level we think the text expresses a single view of a product-service, which we are trying to define. The aim at this level is to determine the wider stance of the author, positive or negative, in a text that includes judgments and opinions (opinionated text). Not recommended for cases in which

PCI '16, November 10 – 12, 2016, Patras, Greece

DOI: http://dx.doi.org/10.1145/3003733.3003769

texts are discussed different aspects of an object or comparing different objects.

Sentence - level: Assignment to sentence-level separates each document proposals whereas each sentence expresses a different emotion. The proposals are identified directly as positive or negative. Moreover, this approach is often associated with the classification of subjectivity (subjectivity classification)

Entity and Features Level: Assignment to Entity and feature level or aspect-based attempts a more detailed analysis in relation to the previous two, and focuses on the same terms and not on an analysis of structural elements of the language (text , sentence, phrase). The opinion holders can give different opinions for different aspects of the same entity like this sentence [4].

2.2 Description of the text-mining methodologies

For export of subjectivity, that is personal opinion - the view for multiple perspectives of a text and of polarity (the degree that is positive, negative or neutral positioning on the theme of a text), two main techniques in the literature that are followed: (a) Lexicon - based and (b) Text Classification or Machine Learning.

2.2.1 Sentiment Analysis based on Lexicon

One of the most common ways of research approach in sentiment analysis field is to create resource dictionaries composed by terms expressing opinion, positive or negative (opinion words) which express emotion and to each is assigned a rating based on the designating percentage classification as positive, negative or neutral (sentiment words). We process these words as independent, ignoring the grammar and the syntactic role or even their position in the text. The expression has been attributed to this approach is "bag of words". Thus, the words are detected by a text analysis which are identical with each sentiment lexicon and the overall feeling of the text is determined by the sum of the scores of individual words and using thresholds (thresholds) in case of multi-level emotional classification.

Such dictionaries are SentiWordNet, Linguistic Inquiry and Word Count, Multi Perspective Question Answering Subjectivity Lexicon (MPQA), General Inquirer etc.

Although the above method is simple and easy to understand and is classified as a popular choice because it is not necessary to make any training, it usually has not high success rates as it takes into account the interaction of words that it sometimes change the semantics of the opinion (negation, words of intensity - Intensifiers, idioms, irony - sarcasm, the finite number of words in dictionaries, emoticons).

2.2.2 Sentiment Analysis with Machine Learning

By the term Machine Learning is defined as a system capable of acquiring and integrating knowledge automatically. In emotional classification of texts, the techniques which applied are divided according to the degree of human intervention in the learning process in [9]:

Supervised Machine Learning: In Supervised Machine Learning, the under education system accepts as input training data, namely examples of data that are already labeled with a feature vector in terms of their emotional content, so that the classifier to recognize and to learn more representative differences between texts belonging to different categories.

Methods Non - Supervised Machine Learning: The system is powered only with inputs invited to discover possible hidden structures between them in order to classify the data into groups that show some similarity. The existence, the number and properties of groups are initially unknown to the system. No training set is required for the extraction of feature vectors. Instead, using pre-built dictionaries of emotion the various terms of the text are characterized and the overall polarity arises.

Semi-Supervised Machine Learning: The under education system is powered with little text data which are labeled as to their emotional content and with enough data which are not labeled.

Although the creation of the dictionary is a time consuming process, this method is considered the most secure and is applied by many researchers because it allows the creation of dictionaries according to user requirements and enables expansion of existing ones.

In our approach we perform opinion analysis on text-level, where the whole user generated comment is analyzed and classified in the proper category and we follow an supervised machine learning methodology where user-generated textual data where annotated by experts.

3. EXPERIMENTAL SETUP

3.1 Data Description

In our research study, we chose to do multi-level categorization of opinion analysis of written texts, thus creating a scale which was: very negative opinion - negative – neutral – positive and very positive. The text approach we followed in our research was to approach the responses of learners with the method of document / text - layer manually giving a score of 1-5 (class label) which accounted overall feeling resulting from the whole answer. Furthermore we used the Entity and Features Level approach by which, any response given commented upon more than one aspect of the same issue and then the corresponding score of 1-5 for each different aspect was annotated by the experts again (Figure 1).

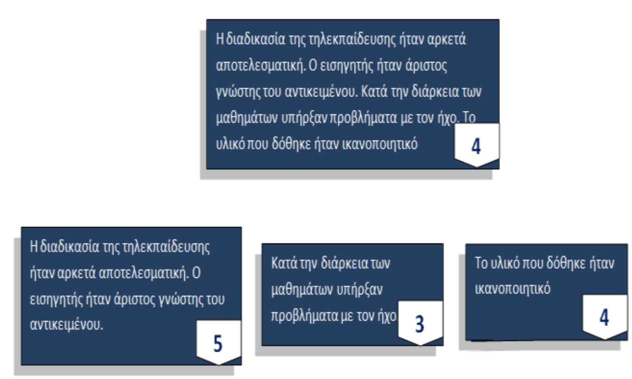

Figure 1: Example of an opinion evaluation and classification in a class label

As shown in the figure below both procedures are followed: (a) during training, we have created a feature extractor and used it to convert each input value to a feature set. We draw these characteristics by the method of text-based features and part-of-speech based features. These feature sets contain information relating to the writing style of each comment. Specifically, features extracted concern how many letters – words- capital letters - small letters -special characters-average word length and digits appear in the user generated comment. Also, information is contained on the feature set relating to what part of speech is each word and how many times it appears in the opinion-based text.

Pairs of feature sets and labels are fed into the machine learning algorithm to generate a model. (b) During prediction, the same feature extractor is used to convert unseen inputs to feature sets. These feature sets are then fed into the model, which generates predicted labels.

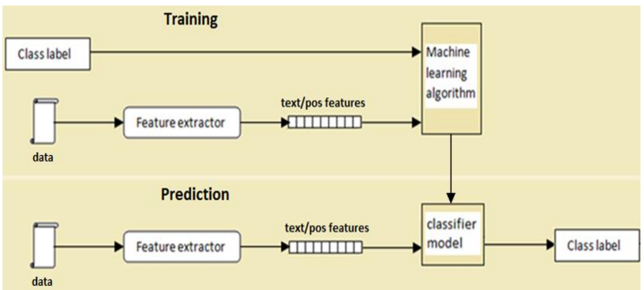

Figure 2: Supervised Classification

With these features extracted for the textual data and the class label annotated by the experts, we train each classification algorithm using a large plurality of data (in the present research to 75%) in order to "understand" the rules based on which feature vectors are classified in the appropriate class label. In this way, the algorithm classifier has the ability to classify new comments represented as feature vectors in the appropriate class (Figure 2).

3.2 Classification

The classification is a technique for data mining, in which an element is assigned to a predetermined set of categories. The term categorization found in the literature also as classification. Generally, the objective of this process is the development of a model, which later can be used for the categorization of future data.

In our work, after the creation of the model, namely the completion of the class label and the pumping of the above features, the next step is its assessment, i.e. the classification process in which the classification algorithms will be examined and compared in terms of their accuracy in sentiment analysis. In the context of our research study were used and examine the performance of algorithms J48, IBk, Multilayer Perceptron, PolyKernel and RBFKernel on the task of data categorization.

4 RESULTS

In our research study we used 11156 user generated comments which constitute the instances of the dataset formulated and each one has been classified in the appropriate class label. Of these, 133 were categorized in class label A (negative opinion), 584 in class label B (unsatisfactory), 3737 in the class C (neutral), 3217 to class D (satisfactory) and 3485 in class label E (very positive).

The graph below shows the overall circumstances which categorized right for Class A for each classification algorithm. Specifically, the best accuracy rate has the J48 with 10 correct classified circumstances while classifiers Multilayer Perceptron_N_500, RBFKernel_C_1 and RBFKernel_C_10 failed to classify correctly any occasion. The accuracy rates for this class were not satisfactory for any classification algorithm in relation to other categories and this is in the fact that there were only 133 occasions to be trained the algorithms before doing the classification.

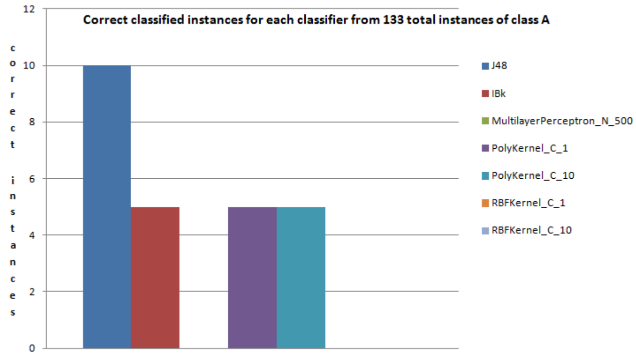

Figure 3: Correct instances of class A for each algorithm

For class B that the number of circumstances are greater than those of class A, precision rates are increased for all algorithms namely the J48 has again the most correct classified circumstances and as shown in (Figure 4) is the most appropriate when we have few occasions for the classifiers to be trained.

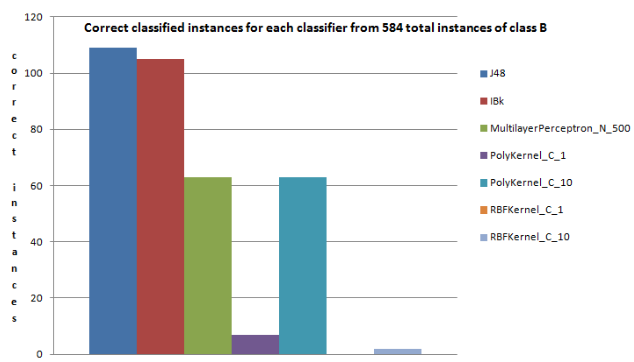

Figure 4: Correct instances of class B for each algorithm

For class C, accuracy rates for all algorithms are very satisfactory, achieving quite good multi-level classification accuracy rate. The greatest accuracy rate is achieved by PolyKernel_C_10 with 63.1% having too close MLP_N_500, RBFKernel_C_10 and J48 Figure 3.

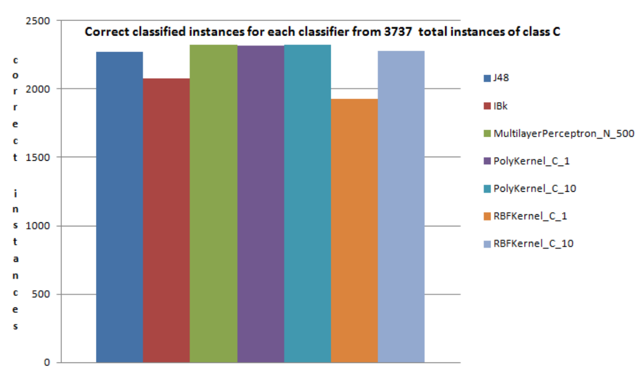

Figure 5: Correct instances of class C for each algorithm

End for classes D & E, precision rates are still quite satisfactory (Figures 6, 7) showing that the model of emotional classification of Greek language, we have created, can have good accuracy rates when we have several occasions to be educated each algorithm.

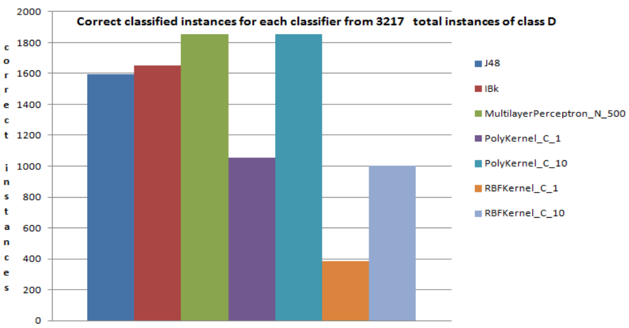

Figure 6: Correct instances of class D for each algorithm

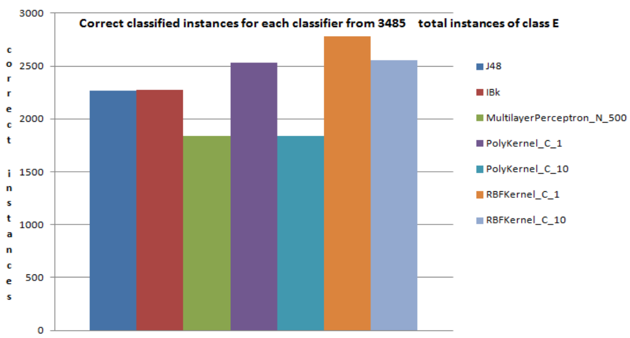

Figure 7: Correct instances of class E for each algorithm

A general assessment of the accuracy (accuracy) achieved in the present research is the fact that for the classes A and B for which the data used to train classifiers were not many in total, 133 and 584 respectively, the success rate was not accordingly well with other classes which were 3737 for class C, 3217 for class D and 3485 data for the class E. For class A, most satisfactory accuracy rate was achieved by IBk 7.5%. For class B, greater accuracy rate noted by the algorithm J48 with 18.5%. For classes C, D and E in which as mentioned above were used fairly to all data on training, there were quite satisfactory success rates. Specifically for the class C the accuracy rates ranging from 51.5% to 63.1% achieved with PolyKernel_C_10. In class D precision rates takes the algorithm MultilayerPerceptron_N_500 with percentage 57.6%. Finally for the class E the best accuracy rate achieves the RBFKernel_C_1 with percentage 79.9%.

Table 1: General Assessment of Accuracy

	J48	IBk	MLP_N_500	PolyKernel_C_1	PolyKernel_C_10	RBF Kernel_C_1	RBF Kernel_C_10
Class A	6%	**7,5%**	3,8%	0%	0%	0%	0%
Class B	**18,7%**	18%	10,8%	1,2%	1%	0%	0.3%
Class C	60,8%	55,7%	62,2%	62%	**63,1%**	51,5%	61%
Class D	49,5%	51,4%	**57,6%**	32,7%	32,8%	12%	31,1%
Class E	65%	65,4%	52,7%	72,7%	72,3%	**79,9%**	73,4%
Average	**56%**	54,9%	54,5%	53%	53,2%	45,7%	52,4%

In summary, the present research created an opinion analysis model of the Greek language based on features that were derived by examining how to write the answer of each student (text-based features) and the display set of each part of speech used in the reply (pos-based features) to multilevel (5 levels) categorization of open-ended response. The accuracy rates achieved for all classifiers that were utilized, was ranging from 45.7% to 56%. The best performance was achieved by the J48 algorithm which reported the highest average accuracy (56%) among the algorithms examined.

5 CONCLUSIONS

In this research, the creation of an opinion categorization model of users' written comments in separate groups through the juxtaposition of the annotations was presented. More specifically, the users' comments in Greek language were classified into five main categories –very negative opinion, negative, neutral, positive and very positive opinion- for a detailed and accurate sentiment breakdown. The proposed framework of this application is quite useful to businesses which desire to monitor the trends of consumers through their views and experiences collocated by the users through the company website or through emails and forum.

As this work is still in progress, there are many directions that future work will examine. First, future work will examine the performance of additional statistical machine techniques and also a larger scale evaluation will be performed in order to provide a more complete insight of the performance of the algorithms. In addition, another interesting aspect that future work will examine concerns the dimensionality reduction of the feature space and the examination of techniques such as feature hashing. Finally, ensemble classification schemas could be also examined and this concerns a main direction of our future work.

6 ACKNOWLEDGMENTS

The related activities that led to these results were performed in the context of the postgraduate MSc Course "Technologies and Infrastructures for Broadband Applications and Services" (http://www.msc.cied.teiwest.gr/mscen/), of the Computer & Informatics Engineering Department – CIED (Technical Educational Institute of Western Greece).

7 REFERENCES

[1] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inform. Retrieval 2 (2008) 1–135.

[2] B. Liu, Sentiment Analysis and Opinion Mining, Morgan and Claypool publishers, 2012 (May)

[3] Cambria, Erik (2016). "Affective computing and sentiment analysis" (PDF). IEEE Intelligent Systems 31 (2): 102–107

[4] *Pang, Bo; Lee, Lillian (2008). "4.1.2 Subjectivity Detection and Opinion Identification". Opinion Mining and Sentiment Analysis. Now Publishers Inc.*

[5] Cambria, E. Schuller, B. Y.-Q. Xia, New avenues in opinion mining and sentiment analysis (extended abstract), in: Proceedings of IJCAI, Buenos Aires 2015.

[6] Pang, Bo; Lee, Lillian (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". Proceedings of the Association for Computational Linguistics (ACL). pp. 115–124

[7] Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.