2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)

Multilingual Sentiment Analysis using Emoticons and Keywords

Georgios S. Solakidis, Konstantinos N. Vavliakis and Pericles A. Mitkas
Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
and
Information Technologies Institute
Centre for Research and Technology Hellas
Thessaloniki, Greece
Email: gsolakid@auth.gr, kvavliak@issel.ee.auth.gr, mitkas@auth.gr

***Abstract*—Nowadays the World Wide Web has evolved into a leading communication channel and information exchange medium. Especially after the introduction of the so-called web 2.0 and the explosion that followed regarding user generated content, the amount of data available over the internet has attracted the interest of both the scientific and business community. Their efforts focus on identifying the inner structures of data and the knowledge that can be derived by analyzing them. Web 2.0 is the subject of study and research in a number of areas. One of these areas is sentiment analysis, where the main goal is to study and draw conclusions about subjectivity, polarity and the feeling that is expressed in user generated content, which mainly consist of free text documents.**

The goal of this paper is to apply sentiment analysis on multilingual data, focusing on documents written in Greek. We developed an integrated framework that accepts user generated documents and then identifies the polarity of the text (neutral, negative or positive) and the sentiment expressed through it (joy, love, anger or sadness).

We followed a semi-supervised approach which led to the development of two techniques for the automatic collection of training data without any human intervention. Our approach involves the detection and use of self-defining features that are available within the data. We take into account two emotionally rich features: a) emoticons and b) lists of emotionally intense keywords. These features are evaluated on data coming from a popular forum, using various classifiers and feature vectors.

Our experimental results point to various conclusions about the effectiveness, advantages and limitations of applying such methods on Greek data. Using keywords we achieved 90% mean accuracy on identifying the subjectivity level and 93% on correctly identifying the polarity level, whereas using emoticons the mean accuracy for each of these levels was 74% and 77% respectively.

***Keywords*: Sentiment Analysis, Greek, Forum, Semi Supervised Learning, Automatic Collection of Training Data, Emoticons, Keywords.**

I. Introduction

Nowadays the World Wide Web has transformed into an interactive ecosystem that allows bidirectional communication. The emergence of blogs, fora and social networking sites that followed the introduction of Web 2.0 has resulted into the accumulation of huge amounts of user-generated data over the web. Sentiment analysis is a research area where the goal is to identify the subjectivity, polarity and emotion expressed in free text documents.

The goal of this work is to design an adaptable and easily extensible framework for the application of document level sentiment analysis on multilingual data, focusing on the Greek language. To this end, we evaluate the advantages and the limitations of using emoticons and emotionally rich keywords in order to collect and tag training data for use in conjunction with common classification algorithms. Moreover, we treat sentiment analysis as a multidimensional problem, putting equal emphasis on a) subjectivity and b) polarity while also adding the level of c) emotion to the final model of the system. Taking these levels of analysis into account, we separate emotion into two disjoint categories: a) positive emotion and b) negative emotion. For the purpose of this study, positive emotion is represented by the sentiments of i) Love, and ii) Joy, while negative emotion includes iii) Anger, and iv) Sadness. As a test-case for evaluating our approach we used anonymized data coming from a popular Greek student forum (www.thmmy.gr) consisting of around 5,600 university students and 1,000,000 posts. Four are the main contributions of our work, namely:

1) We apply sentiment analysis, including subjectivity, polarity and emotion classification on Greek data using language independent techniques.
2) We compare the ability of emoticons and keywords to define the text in which they are included.
3) We do not include emoticons as features in any of our sets in order to test whether they are used as sentiment intensifiers or sentiment declarants.
4) We treat the emotion level as an extension to the polarity level of analysis, separating it into two distinct categories including positive and negative emotion.

The paper is organized as follows: Section II presents related work, while Section III provides an overview of the proposed system and discusses the various submodules. Section IV outlines the methods used for automatic tagging of the data used for training. Section V describes the proposed analysis focusing on keywords and emoticons, and then Section VI

978-1-4799-4143-8/14 $31.00 © 2014 IEEE
DOI 10.1109/WI-IAT.2014.86

compares the performance of the two elements. Section VII presents the final model and the results of evaluation with real data, while Section VIII presents additional experiments. Finally, Section IX summarizes our conclusions and discusses future work.

II. Related Work

Sentiment analysis is a fascinating problem and a research area where a number of different approaches have been proposed. One popular approach makes use of lexicons with sentiment-related term weights [1], while another approach makes use of machine learning [2], a process that involves training classifiers that identify the sentiment in free text documents.

A major disadvantage of machine learning methods for sentiment classification is the need of human supervision in the process of training set creation. Manual annotation of data is a tedious task which puts a limit on the resulting training sets in terms of size and variety. One way to overcome this limitation is the adoption of multidisciplinary approaches, such as active learning. In this case, a small portion of annotated data is used in order to classify unknown data which is then incorporated into the training set [3]. Another approach involves finding self-defining elements within the data in order to automatically form the training sets. Elements like emoticons and keywords have been studied in the past [4], [5], [6], [7], yielding promising results.

In this paper we use a machine learning approach relying on the automatic collection of training data using, evaluating and comparing the results of two separate elements: a) emoticons and b) keywords. Emoticons have been studied in the past as visual cues bearing great significance in terms of their ability to convey the sentiment expressed or even implied in a given text. They have been used both in lexicon-based approaches [4] and machine learning approaches [5]. Hogenboom et al. [4] conclude that users usually use emoticons in order to express or disambiguate sentiment in particular text segments, thus rendering them potentially better proxies for sentiment identification than textual cues. Applying their lexicon-based approach on 2,080 Dutch tweets and forum messages they achieved accuracy ranging from 59% (sentence level) to 94% (paragraph level). Keywords have also been studied as a combined approach using lexicon-based and machine learning-based methods in order to classify sentiment [6], [7]. Zhang et al. [6] used automatically formed training sets in combination with an augmented lexicon in order to classify the polarity of tweets, achieving an average accuracy of 85.4%.

Regarding machine learning approaches, Pang et al. [2] test various algorithms and feature vectors on datasets consisting of movie reviews in order to evaluate their performance. One of their conclusions was that the use of simple machine learning algorithms and simple feature vector representations are capable of achieving satisfying baseline performance. Our experiments confirm this conclusion.

In the past, sentiment analysis has been studied both as a single-dimensional and a multi-dimensional problem. There have been single-dimensional approaches that took into account only the polarity level [2], [8], the subjectivity level [9] or the emotion level of analysis [10]. Other approaches took into account both the levels of subjectivity and polarity [11], or the levels of subjectivity and emotion [10]. Although both the levels of subjectivity and polarity constitute binary classification tasks, analysis on the level of emotion can be treated as a 2-class or a multi-class problem, depending on the number of emotions used. Mihalcea et al. [10] classified documents into two categories: those denoting happiness and those denoting unhappiness. On the other hand, Balabantaray et al. [10] took into account all six basic emotions as defined by Ekman [12], with the addition of the neutral class in order to integrate the level of subjectivity into their research. Our approach takes into account all three levels of analysis, treating emotion as an extension to the level of polarity. Instead of considering emotion as a multi-class problem, we separate it into two distinct categories: a) positive emotion and b) negative emotion, thus reducing an otherwise 4-class emotion analysis task to a binary classification problem. Finally, sentiment analysis is also a problem tightly related to natural language processing and as such it has been studied in a variety of different languages, including English, Chinese, Hindi etc. [13], [14]. Our research is focused on Greek data, therefore extending this list.

III. System Overview

The proposed system integrates and automates all the tasks associated with semi-supervised emotion detection, including a) automatic collection and tagging of the data, b) automatic formation of the training sets, c) preprocessing of the data, d) training of the classifiers, e) model evaluation, and f) application of sentiment analysis. By implementing each task as a separate module, we improved system accessibility and scalability. As a result, the proposed framework consists of 9 subsystems, as depicted in Figure 1:

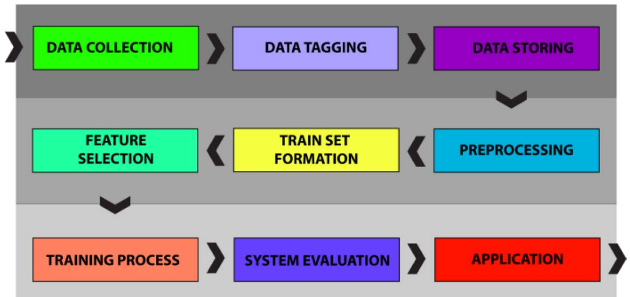

Fig. 1. System Overview

These subsystems include:

- Data collection: Automatic collection of public user posts using web scraping. A total of 750,000 posts were collected from the thmmy.gr forum. In order to maintain user anonymity no usernames were collected.

- Data tagging: A set of heuristic rules putting tags on each post. Tags refer to whether the post is considered subjective, neutral, positive or negative according to the emoticons and the list of keywords used in our research.
- Data storing: Storing and indexing the tagged data.
- Data preprocessing: The tasks implemented by this subsystem include: a) the transformation of all Greek characters into their respective Latin representation, b) removal of all punctuation marks, c) capitalization of all letters, d) removal of all special characters, e) stemming, f) removal of all repeated characters, g) removal of Greek stop words, h) removal of all posts the length of which exceeded 300 characters. The last step can be justified by the fact that long posts usually provide information or express opinion rather than sentiment.
- Creation of training sets: Twelve training sets were created. Since comparing emoticons and keywords refers to the levels of a) subjectivity and b) polarity, the training sets for the level of emotion were created using the keywords method only.
- Feature selection: In our experiments we tested a number of different feature vectors, including a) Unigrams, b) Unigrams and the top 200 Bigrams, c) Unigrams and the top 200 Trigrams, and d) Unigrams, the top 200 Bigrams, and the top 200 Trigrams. For all feature vectors we only took into account their presence or absence in a given post (binary representation), and we used a) all, as well as the best b) 2,000, c) 4,000, d) 6,000, e) 8,000 and f) 10,000 unigrams. The algorithm for selecting the best features is presented in Listing 1:

```
for each word feature do:
  count_0=Frequency_of_appearance_
            in_class_0
  count_1=Frequency_of_appearance_
            in_class_1
end for
for each word feature do:
  score_0= Associate(class_0_word_count,
           total_word_count,count_0)
  score_1= Associate(class_1_word_count,
           total_word_count,count_1)
end for
for each word feature do:
  word_score= score_0 + score_1
end for
best_words= sort_words(word_scores)
return n words from best_words
```

Listing 1. Code for best feature selection

- Training process: For our experiments we tested 5 classification algorithms including a) Multinomial Naive Bayes (MUL), b) Bernoulli Naive Bayes (BER), c) Logistic Regression (LOG), d) Linear SVM (SVM), e) K-Nearest Neighbors (KNN).
- Model Evaluation: Our models were evaluated using a) 5-fold cross validation, and b) an external test set consisting of i) 2,000 neutral posts and ii) 2,000 subjective posts for the level of subjectivity and iii) 1,000 positive and iv) 1,000 negative posts for the level of polarity. The level of emotion was tested using only 5-fold cross validation. The user posts included in the external test set were evaluated manually by 2 reviewers until they reached agreement.
- Application: After the evaluation of keywords and emoticons, the winning element along with the classifier and the feature vector that yielded the best results formed the final model of the system. This model was used to classify previously unseen real data collected from the same forum as well as other external sources, a process which provided insights about the behavior of our model.

IV. Training Set Formation

A. Emoticons

First, we created a list of emoticons available in our dataset. Next, we separated these emoticons into three categories: a) positive, b) negative and c) undefined. The first two categories consist of emoticons that clearly express either positive or negative sentiment, based on their characteristics, their unique name, as well as early experiments. The third category consists of emoticons whose polarity is unclear. Fourteen emoticons were used for the positive and negative polarity. After that, we developed a tagging subsystem which was used to automatically tag user posts as positive, negative or none depending on the presence (positive, negative) or the absence (none) of emoticons in a specific post. Since more than one emoticons can be used in each post, we applied the following rules:

- If a post contains more than one emoticon of the same polarity, then the post is considered valid and it can be used in the training process.
- If a post contains more than one emoticon of different polarities, then the post is considered invalid and it cannot be used in the training process.

The above process resulted in the creation of three training sets a) positive, b) negative and c) subjective set, consisting of both positive and negative posts. Emoticons were not included as features in any set, in order to test whether emoticons are used by users as a means to intensify a sentiment already expressed in textual form.

B. Keywords

First, we created a list of 30 emotionally intense positive and negative words, such us *love, despise, hate* etc. After applying stemming, we automatically gathered all user posts containing any of these words in any form, derivatives included. This process resulted in the creation of the training sets for the positive and negative polarity, the combination of which formed the training set for subjectivity. Using keywords, we were able to tag a total of around 25,000 user posts. In order to compare the potential effectiveness of keywords and emoticons, we decided to use training sets of equal sizes for both methods. Thus, the training sets for the positive and negative polarity comprised of 4,850 randomly selected user posts, while the subjectivity train set obtained of a total of 5,000 posts.

The training set for the neutral class is common for both methods. 2,000 user posts were included in this set, which were evaluated by 2 reviewers until they reached agreement.

C. Training sets for the level of emotion

The training sets for the level of emotion were created using the method of keywords, which gave the best results during the phase of evaluation. For this reason, the above training sets were created by separating the positive and negative user posts into four categories depending on how relative they were to the sentiments of love, joy, anger, and sadness. To this end, the tagging process was repeated exclusively for the level of emotion. The final training sets for each category included 2,000 posts for love, 2,000 posts for joy, 1,000 posts for anger and 1,000 posts for sadness.

D. Set size and dependencies

The dependencies among the different training sets are depicted in Figure 2.

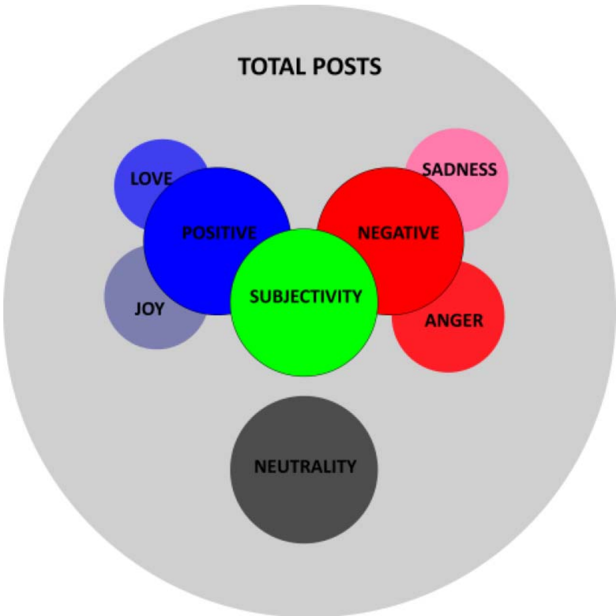

Fig. 2. Dependencies of the various train sets

The formation of the training sets begun with identifying the level of polarity. After automatically separating positive and negative posts, a randomly generated subset combining both the positive and negative training sets was used for identifying the level of subjectivity. Moreover, the keywords used to gather positive and negative posts were manually separated into four categories, according to how closely related they were to the emotions of a) love, b) joy, c) anger and d) sadness. The resulting lists were used in order to automatically form the training sets for the 4 emotions. Finally, the training set for neutrality was also automatically created by collecting posts from topics related to neutral entities. The neutrality set was the only training set to be manually evaluated after its formation.

Table I presents the size of each training set for both methods (emoticons, keywords):

TABLE I
SIZE OF THE TRAINING SETS.

Emoticons		
Classification Pair	Class 0	Class 1
Subjective-Objective	5,000	5,000
Positive-Negative	4,850	4,850
Keywords		
Subjective-Objective	5,000	5,000
Positive-Negative	4,850	4,850
Love-Joy	2,000	2,000
Anger-Sadness	1,000	1,000

V. TRAINING PROCESS

During the training phase, all training sets were used in order to find the best combination of classifiers and feature vectors for both elements (emoticons, keywords) on the levels of subjectivity and polarity. Since we tested 5 classifiers using 4 different feature vectors with each containing 6 different numbers of unigrams, a total of 240 models were trained for the method of keywords and another 240 models for the method of emoticons. These models were first evaluated using 5-fold cross-validation with 20% of the posts used for testing. Following this, the best models in terms of precision, recall, F-measure and accuracy were reevaluated using the external test sets. In order to test the versatility of both elements, our next test involved using the training sets of each element as testing sets for the other. This gave us insight into the behavior of keywords when tested on user posts containing emoticons and the behavior of emoticons when tested on posts lacking this element. Finally, we computed the mean scores in terms of accuracy for each of these models. The model with the highest overall accuracy was chosen to represent each element on the level of subjectivity and polarity. K-Nearest Neighbors was not included in the following tables since it was the classifier with the lowest overall performance.

A. Keywords

For the level of subjectivity, the best overall performance was achieved by the Logistic Regression classifier (LOG) and the feature vector containing the best 2,000 unigrams. Table II shows the mean accuracy scores of the best models:

TABLE II
BEST MODELS FOR THE LEVEL OF SUBJECTIVITY - KEYWORDS.

Classifier	MUL	BER	SVM	LOG
Cross Validation	0.877	0.895	0.942	0.936
Test Set	0.772	0.826	0.865	0.872
Mean Accuracy	0.824	0.860	0.903	0.904

For the level of polarity, the best overall performance was achieved by the Logistic Regression classifier and the feature vector containing the best 6,000 unigrams along with the best

200 bigrams. Table III presents the mean accuracy of the best models:

TABLE III
BEST MODELS FOR THE LEVEL OF POLARITY - KEYWORDS.

Classifier	MUL	BER	SVM	LOG
Cross Validation	0.963	0.955	0.970	0.968
Test Set	0.882	0864	0.892	0.897
Mean Accuracy	0.922	0.909	0.931	0.932

B. Emoticons

For the level of subjectivity, the best overall performance was achieved by the Multinomial Naive Bayes classifier (MUL) and the feature vector containing the best 4,000 unigrams. The results are presented in Table IV.

TABLE IV
BEST MODELS FOR THE LEVEL OF SUBJECTIVITY - EMOTICONS.

Classifier	MUL	BER	SVM	LOG
Cross Validation	0.808	0.796	0.794	0.796
Test Set	0.674	0.579	0.628	0.629
Mean Accuracy	0.741	0.687	0.711	0.712

For the level of polarity, the best overall performance was achieved by the Multinomial Naive Bayes classifier and the feature vector containing the best 8,000 unigrams. Table V presents the mean accuracy of the best models:

TABLE V
BEST MODELS FOR THE LEVEL OF SUBJECTIVITY - EMOTICONS.

Classifier	MUL	BER	SVM	LOG
Cross Validation	0.785	0.764	0.768	0.766
Test Set	0.754	0.718	0.763	0.766
Mean Accuracy	0.769	0.741	0.765	0.766

VI. COMPARING EMOTICONS-KEYWORDS

In this part we compare the models selected to represent the two elements on the levels of subjectivity and polarity.

A. Subjectivity level

According to Figure 3, the element of keywords outperforms that of emoticons on all measures. However, both elements achieve scores higher than 70%, which is indicative of their ability to define the text in which they are included.

B. Polarity level

As depicted in Figure 4, the overall performance of both elements is slightly increased on this level, with the element of keywords achieving significantly better scores.

C. Testing keywords using the training sets of emoticons

Figure 5 presents the performance of keywords on both the subjectivity and the polarity levels when tested on user posts containing emoticons.

In Figure 5 a significant drop in the overall performance of this element is presented. Specifically, considering the fact that the precision of keywords was reduced by almost 20% (compared to Figures 3 and 4) we assume that the training sets of emoticons contain posts in which the sentiment expressed by words is contradictory to the sentiment representing the class assigned to them. A qualitative analysis of the training data labeled by emoticons confirms this assumption, revealing that a) many posts labeled as subjective contained no subjective features at all, and b) posts labeled as positive (negative) contained features typically found in the negative (positive) class. Moreover, the reduction of the recall by almost 30% (compared to Figures 3 and 4) reveals that the datasets of emoticons contain features unknown to the training sets of

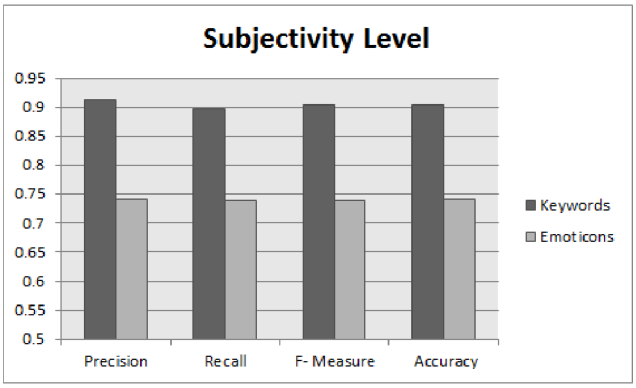

Fig. 3. Comparison: subjectivity level

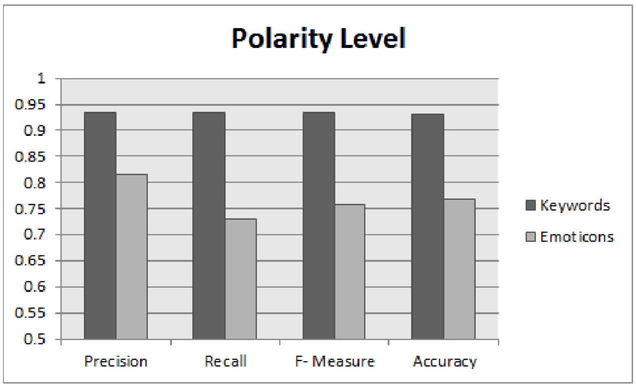

Fig. 4. Comparison: polarity level

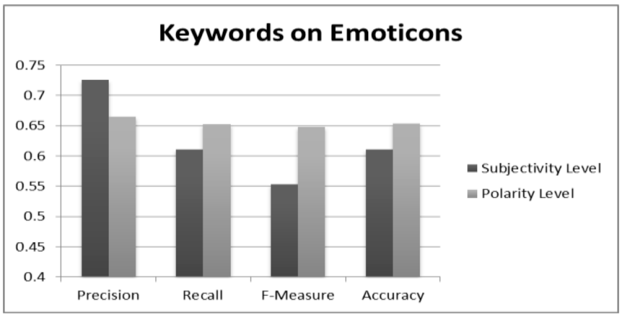

Fig. 5. Keywords tested using the emoticons as training set

keywords, indicating that: i) using emoticons it is possible to gather a wide variety of different words and expressions regarding subjectivity and polarity, and ii) using only a small list of keywords it is impossible to gather all the sentiment-indicating expressions used.

D. Testing emoticons using the training sets of keywords

Figure 6 presents the performance of emoticons on both the subjectivity and the polarity levels when tested on user posts that may or may not contain this element.

Despite the results of the qualitative analysis presented earlier and despite the fact that the training sets of emoticons contain a number of posts bearing dubious meaning, the scores achieved in this test indicate that although in many cases emoticons are used as a means to imply or disambiguate the text in which they are included, in most cases they simply act as intensifiers, which justifies the accuracy scores of this element ranging from 75% to 80%, as depicted in Figures 3, 4 and 6.

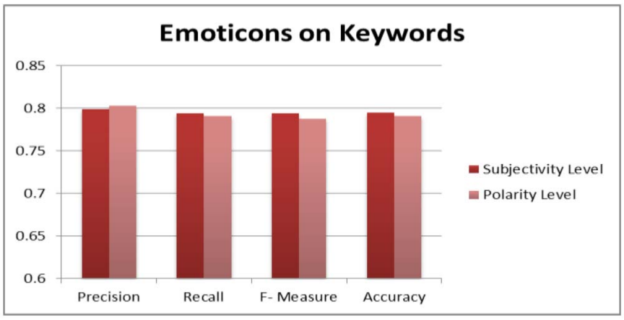

Fig. 6. Emoticons tested using the train sets of keywords

E. Comparison on both levels

In this test, the models representing each element (emoticons, keywords) were tested using a dataset formed by combining the testing sets used for the levels of subjectivity and polarity. The resulting testing set consisted of 2,000 subjective, 1,000 positive and 1,000 negative posts. The purpose of this test was to evaluate the accuracy of keywords and emoticons using a 2-level system that classifies input into 3 different classes, the 1) neutral, 2) positive and 3) negative class. The results are depicted in Figure 7, where we can see that keywords achieve an accuracy of 86%, outperforming emoticons by almost 25%.

VII. System Model and Application

The system model consists of the models trained using the keywords method. The models for the level of emotion were trained using the Logistic Regression classifier with the feature vector containing the best 2,000 unigrams for both the positive and the negative sentiment respectively. Using 5-fold cross-validation, the accuracy for the level of emotion was 96% for the positive emotion (love, joy) and 97% for the negative emotion (anger, sadness). These results indicate that

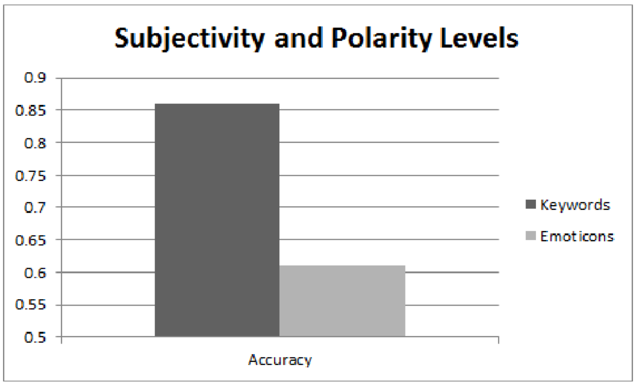

Fig. 7. Comparison on both levels

based on the keywords used, our system is able to distinguish the emotion of love from the emotion of joy and the emotion of anger from the emotion of sadness. Figure 8 presents the results for both the positive and negative emotion classes.

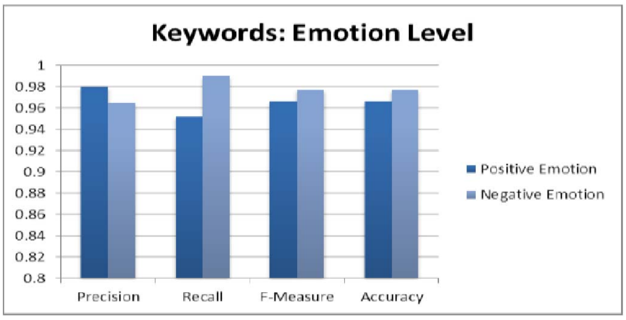

Fig. 8. Results for the level of emotion

Table VI depicts the classifiers, feature vectors and number of best features used in each level of analysis. The element of keywords was chosen as the element used in the final model of our system due to its overall superior performance (Section V).

TABLE VI
System Model

System Classification Model	
Winning Element	Keywords
Levels of Analysis	3
Subjectivity Level	
Classifier	Logistic Regression
Feature Vector	Unigrams
Features	2,000
Polarity Level	
Classifier	Logistic Regression
Feature Vector	Unigrams + Bigrams
Features	6,000 + 200
Emotion Level	
Classifier	Logistic Regression
Feature Vector	Unigrams
Features	2,000

The integrated model was used to classify previously unseen real user posts which were collected automatically using various keywords in the sense of entities. Finally, the model was used to classify all 750,000 posts collected from the forum. Neutrality was detected in 88% of the user posts, leaving only a 12% of subjective user posts. Polarity analysis showed that 52% of the subjective user posts were positive. Finally, emotion analysis showed that love is expressed in 9%, joy in 43%, anger in 46% and sadness in only 2% of the total posts. Additionally, 434,592 emoticons of any kind were used in 30% of the total users posts.

VIII. Additional Experiments

In this section we present additional experiments in order to evaluate certain aspects of the methods used in our approach. Specifically, we evaluate a) the importance of stemming and b) the combination of different classifiers using i) simple voting and ii) voting based on the maximum probability rule.

A. The importance of stemming

Greek is a complex language that features a large number of grammatical rules. Considering the fact that homogenization of the data would be a crucial part of our research, stemming was used in all feature vectors in order to bypass the restrictions posed by those rules. Figure 9 presents the effect of stemming on the performance of the chosen model. For this experiment we used the model of the keywords method for the level of subjectivity (Table VI).

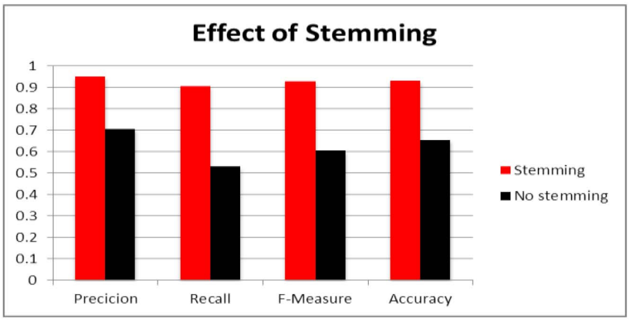

Fig. 9. The importance of stemming

According to the results of this experiment, when stemming is not used the performance of the model is reduced by almost 30% percent on all measures.

B. Combination of different classifiers

In this experiment we use two voting schemes in order to evaluate the performance of the combination of the best models according to Section V (Tables II and III). First, we used a simple voting scheme in which a post was classified according to the majority rule. In Figure 10, we present the results of this experiment, for which we used the models trained for the level of polarity by a) Logistic Regression, b) Multinomial Naive Bayes and c) Linear SVM.

According to the results of this experiment, combining different classifiers under a simple voting scheme slightly

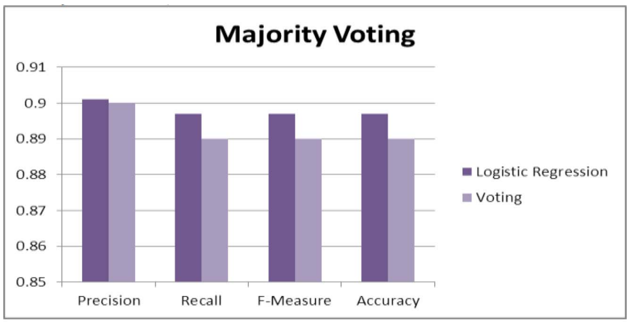

Fig. 10. Voting scheme based on majority rule

reduces performance. Before evaluating the maximum probability scheme, we repeat this experiment for the level of subjectivity, using a) Logistic Regression, b) Bernoulli Naive Bayes, c) Linear SVM. The results are presented in Figure 11.

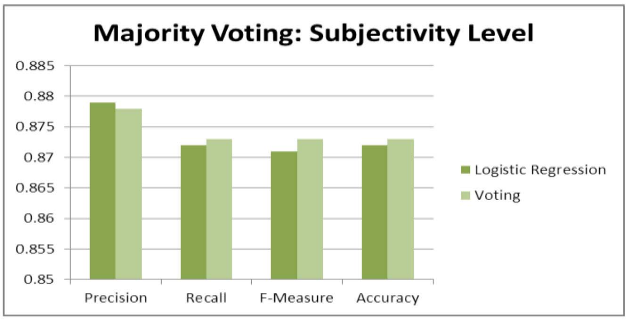

Fig. 11. Voting scheme based on majority rule: subjectivity level

According to Figure 11 the use of simple voting slightly increases the overall performance on the level of subjectivity. Taking into account the effect of this scheme on both levels, it is clear that performance does not differentiate enough to either use or discard it. In the final experiment we use a voting scheme based on the maximum posterior probability of two classifiers. Figure 12 presents the results of applying this scheme on the level of subjectivity. The classifiers used are a) Logistic Regression and b) Bernoulli Naive Bayes. As depicted in Figure 12, applying this scheme has a negative effect, reducing performance by 2.5% to 3%.

IX. Conclusions and Future Work

A. Greek data and general approach

Applying sentiment analysis on Greek data is a daunting problem considering the overall complexity of the Greek language and especially the plethora of syntactical and grammatical rules. Taking into consideration the lack of Greek bibliography regarding the topic of sentiment analysis, we decided to develop a system covering all the major aspects of analysis, putting emphasis on the process of homogenizing the data (stemming, use of a Latin representation) while also relying on it to bring out patterns hidden in its structures. Performing various experiments, we tested the effect of using

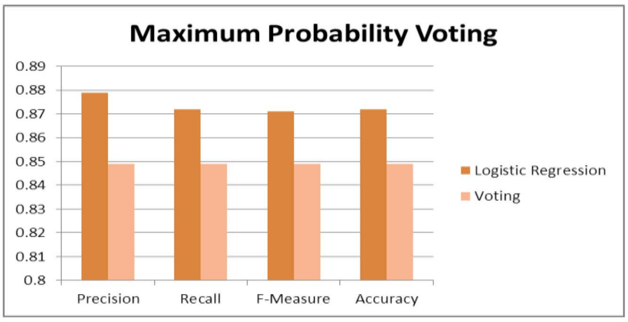

Fig. 12. Voting scheme based on maximum probability

simple methods (single classifier) over more complex methods (voting) and simple feature vectors (unigrams) over vectors containing bigrams and trigrams. Our results, which include accuracy scores ranging from 65% to 93% (Section V) lead us to the conclusion that using elements such as emoticons and keywords in order to automatically form training sets in conjunction with a single classifier and a simple bag of words representation is an effective approach.

B. Emoticons

In this work, we evaluated the role and the importance of emoticons by treating them as features capable of labeling the text in which they are included. Despite their comparatively lower performance (Sections V and VI), taking into account that emoticons were present in 30% of the total user posts used in our research as well as the ineffectiveness of the keywords method to maintain its high performance when facing data marked by this element, we conclude that emoticons are significant features that are used to both intensify and declare the presence and the polarity of a specific sentiment within a document. Moreover, qualitative analysis of our data confirms that emoticons are often used to imply or disambiguate sentiments [4], a fact that is partly responsible for the low levels of precision and recall presented in Section V. In order to support this conclusion, our tests involved purposely removing emoticons from posts after the formation of the training sets, thus highlighting the semantic relationship between the textual and visual cues used by users in order to express sentiment.

C. Keywords

Despite using only a small number of keywords to collect the training data, this element was able to outperform emoticons on all the different measures. While a significant drop in the overall recall of the various models is to be expected when it comes to real data, this element is capable of achieving significant results especially if combined with lexicons in order to counterbalance the lack of keyword variety. However, the method of keywords is able to identify sentiments that are clearly expressed in a particular text segment, a limitation that can be overcome by developing a combinational approach taking advantage of both emoticons and keywords.

D. Future Work

Future work includes a) the development and evaluation of a combinational approach taking advantage of both emoticons and keywords, b) use and evaluation of feature vectors containing specific pos-tags such as adjectives and adverbs [8], [15], c) use and evaluation of feature vectors containing specific punctuation marks and special characters, and d) further testing our approach in a multilingual context.

References

[1] Y. H. Cho and K. J. Lee, "Automatic affect recognition using natural language processing techniques and manually built affect lexicon," *IEICE - Trans. Inf. Syst.*, vol. E89-D, no. 12, pp. 2964–2971, Dec. 2006.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, ser. EMNLP '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 79–86.

[3] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '03, 2003, pp. 105–112.

[4] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting emoticons in sentiment analysis," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, ser. SAC '13. New York, NY, USA: ACM, 2013, pp. 703–710.

[5] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL Student Research Workshop*, ser. ACLstudent '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 43–48.

[6] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexiconbased and learning-based methods for twitter sentiment analysis," *HP Laboratories, Technical Report HPL-2011*, vol. 89, 2011.

[7] S. Tan, Y. Wang, and X. Cheng, "Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 743–744.

[8] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 417–424.

[9] J. Wiebe, "Learning subjective adjectives from corpora," in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press, 2000, pp. 735–740.

[10] R. C, M. Mohammad, and N. Sharma, "Article: Multi-class twitter emotion classification: A new approach," *International Journal of Applied Information Systems*, vol. 4, no. 1, pp. 48–53, September 2012, published by Foundation of Computer Science, New York, USA.

[11] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 129–136.

[12] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[13] B. Lu, B. K. TSOU, and T. JIANG, "Supervised approaches and dependency parsing for chinese opinion analysis at ntcir-8," *Proceedings of NTCIR-8 Workshop Meeting. Tokyo, Japón*, 2010.

[14] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 976–983.

[15] V. Hatzivassiloglou and J. M. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, ser. COLING '00. Stroudsburg, PA, USA: Association for Computational Linguistics, 2000, pp. 299–305.