# Mining and estimating users' opinion strength in forum texts regarding governmental decisions

George Stylios [1], Dimitrios Tsolis[2] and Dimitrios Christodoulakis[2]
[1] Technical Educational Institute of Ionian Islands, Dept. of Applications of Information Technology in Administration & Economy, 3100, Lefkas, Greece.
[2] University of Patras, Computer Engineering & Informatics Dep., 26504, Patras, Greece
gstylios@teiion.gr, dtsolis@upatras.gr, dxri@ceid.upatras.gr

**Abstract.** Web 2.0 has facilitated interactive information sharing on the WWW, allowing users the opportunity to articulate their opinions on different topics. In this framework, certain practices implement information monitoring systems so as digests, reports on keywords and thematic queries regarding opinions on government decisions to be created. Analysis of rubrics associations, primary semantic and statistical interpretation of the texts is usually carried out. It is, on the other hand, rather difficult to get punctual predicts and estimate sufficiently forum users' opinion strength. In this work we present a methodology which automatically mines and estimates the strength of users' opinions on text forums regarding government decisions. According to our methodology, quantitative features are automatically mined from forum posts and then passed to a Support Vector Machine based classifier where the users' opinion strength is estimated. The proposed methodology has been validated in real data and initial experimental results are presented.

**Keywords:** Opinion mining, opinion strength mining, sentiment analysis, E-Government, Knowledge extraction, linguistic analysis, Machine learning, Support Vector Machines.

## 1 Introduction

It is well known that "What other people think" has always been an important piece of information during the decision making process. Long before awareness of the World Wide Web became widespread, many of us asked our friends who they were planning to vote for in local elections, requested reference letters regarding job applicants from colleagues, or consulted Consumer Reports to decide what product to buy. But the internet and the Web have now (among other things) made it possible to find out about the opinions and the experiences of those in the vast pool that are neither our personal acquaintances nor well known professional critics – that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the internet [10]. On the other hand, businesses have used data mining, for years, to analyze customer demographics and transaction history to better target direct marketing efforts. Recent advances in computer speed and the collecting data by many businesses have inspired the improvement of software to

achieve today's mining abilities. As parallel processing and the use of artificial intelligence have met with improvements in software and growing business awareness of the benefits of database analysis, DM and related fields, based on both statistical tools and computer science, have emerged. In addition, wikis, social networking and folksonomies are often focused on personal life, and many on professional life. Web 2.0 enhances the creativity, collaboration, information sharing and functionality of the web. In the professional or business environment, both private and public sectors are very interested in offering the best services to the users [2]. With the explosion of the Web 2.0 platforms such as blogs, discussion forums, peer to peer networks, and various other types of social media citizens have at their disposal a soapbox of unprecedented reach and power by which to share their experiences and opinions positive or negative, regarding any product or service [20].

Opinion mining has recently become a topic of interest trying to combine statistics, Artificial Intelligence and Data Mining technologies in a unified framework [10]. Negative and positive opinions can be used as guidelines for companies to change their strategies toward specific target groups, customers to decide on the purchase of a product or destination place for their holidays and lately for governments to improve services, launch campaigns etc [9]. Traditionally the opinion of the people was acquired through Gallup polls and questionnaires. The latest trend however is to extract public opinion expressed in text documents in the web (blogs, forums), information that might be more objective since it is expressed without any "pressure". On the other hand the tendency of a person for or against an argument, a product etc is not as easily extracted as in the case of specific questionnaires. It is therefore somewhat subjective posing an extra difficulty in the analysis of this information.

Opinion mining is an uprising technology also in Electronic government. E-government is a way for governments to use the most innovative information and communication technologies to offer citizens efficient access to information and services [7]. E-government, is correlated with the use of digital technology in the management and delivery of public services, by enhancing the efficiency of the public sector and developing more personal, customized relations between citizens and their government. The Semantic Web plays a crucial role in automatic delivery of customized e-government services. It extends the existing Web by providing a framework for technologies that give meaning to data and applications for automatic processing [4]. In addition, Web 2.0 plays an important role in the opinion sharing, voting and open discussions of citizens in crucial governmental decisions. Opinion mining offers a solid basis for new citizen oriented e-government services

## 2 Related Work

As is well known, opinions matter a great deal in politics. Some work has focused on understanding what voters are thinking, whereas other projects have as a long term goal the clarification of politicians' positions, such as what public figures support or oppose, to enhance the quality of information that voters have access to. The field of

web opinion mining and sentiment analysis is well-suited to various types of intelligence applications e.g. Government intelligent. Web opinion mining aims to extract, summarize, and track various aspects of subjective information on the Web. Ku [9], applied web mining techniques to mine positive and negative sentiment words and their weights on the basis of Chinese word structures. Xu [19] proposed a system for opinion mining using poll results on the web dealing with opinion answering question, opinion mining on a single object and opinion mining on multiple objects. Furuse [3] developed a search engine that can extract opinion sentences relevant to an open-domain query -based not only on positive or negative measurements but also on neutral opinions, requests, advice, and thoughts- from Japanese blog pages. Miao [12] proposed AMAZING a sentiment mining and retrieval system which mines knowledge from consumer product reviews by utilizing data mining and information retrieval technology based on a ranking mechanism taking temporal opinion quality and relevance into account to meet customers' information needs. Zhai [21] developed Opinion Observe to compare consumer opinions of different products based on online reviews, while Sun [8] created BlogHarvest which is a blog mining and search framework that extracts the interests of the blogger, finds and recommends blogs with similar topics and provides blog oriented search functionality.

An opinion utility named Jodange was built in the Leveraging Cornell University. Jodgane identifies opinion holders on issues, organizations, or people of interest. It can track the impact of an issue via publication, region, opinion holder, tonality or any other measurement, uncover important sentiment trends on key issues and correlate opinions against specific outcomes. VIStology's IBlogs (http://www.vistology.com/about/about.html) project, funded by the Air Force Office of Scientific Research's Distributed Intelligence provides blog analysts a tool for monitoring, evaluating, and anticipating the impact of blogs by clustering posts by news event and ranking their significance by relevance, timeliness, specificity and credibility, as measured by novel metrics. This technology allows analysts to discover, from the bottom up, the issues that are important in a local blogosphere, by providing measurements particular to that locale alone. The need for identifying opinions has motivated a number of automated methods for detecting opinions or other subjective text passages [15], [16], [17], [6] and assigning them to subcategories such as positive and negative opinions [10], [13], [18]. A variety of machine learning techniques have been employed for this purpose generally based on lexical cues associated with opinions. However, a common element of current approaches is their focus on either an entire document [10], [13] or on full sentences [16], [6].

Although all the above mentioned research deals with web opinion extraction, according to our knowledge there is no previous work reported regarding automated assessment of blog or post user's opinion strength. Apparently, it is of great importance not only to extract someone's opinion (positive or negative), but also to estimate if someone supports his/hers opinion with arguments or epicheiremas (i.e opinion strength). In the following sections, initially we describe our methodology which automatically estimates post/blog users' opinion strength. The proposed methodology is validated using real data coming from the website of a newspaper. The initial results are provided next in the experimental evaluation section. Finally, our conclusions and future work are described in the concluding remarks section.

## 3 The basis of the proposed methodology

Automatically determining posts provided from users using arguments to support their personal opinion (positive opinion strength) would help in selecting the appropriate type of information given an application and in organizing and presenting that information. Text materials from many web sources (e.g., posts, blogs) usually mix facts and opinions. In this work, we provide a methodology that automatically classifies user's opinion strength into two classes, high or low, using quantitative features being extracted from posts or blogs. For that reason a Support Vector Machine (SVM) classifier is employed [14]. A classification task based on SVM usually involves training and testing data, which consist of a number of data instances. Each instance in the training set contains one ''target value'' (class labels) and several ''attributes''. The goal of the SVM is to produce a model, that predicts a target value of data instances in the testing set in which only the attributes are given. Let a training set of instance-label pairs be

$$(\text{xi, yi}),\ \text{i} = 1, \ldots, \text{p} \tag{1}$$

where $x_i$ is the training vector of original data belonging to one of two classes (high opinion strength, low opinion strength), p is the number of the blogs/posts, $y_i \in \{-1,1\}$ indicates the (one of the two) class of $x_i$. The support vector machine requires the solution of the following optimization problem:

$$min_{w,b,\xi}\left(\frac{1}{2}w^T w + c\sum_{i=1}^{p}\xi_i\right), \tag{2}$$

subject to

$$y_i(w^T\varphi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{3}$$

where b is the bias term, w is a vector perpendicular to the hyperplane ⟨w, b⟩, ξ the factor of classification error and c>0 is the penalty on parameter of the error term. The training vectors $x_i$ are mapped into a higher dimensional space F by the function $\varphi: R^n \rightarrow F$, where F is a feature space where the data are separable. SVM finds a separating hyperplane with the maximal geometric margin and minimal empirical risk $R_{emp}$ in this higher dimensional space. $R_{emp}$ is defined as

$$\mathcal{R}_{emp}(a) = \frac{1}{2p}\sum_{i=1}^{p}|y_i - f(x_i, a)| \tag{4}$$

where f is the decision function defined as

$$f(x) = \sum_{i=1}^{p} y_i a_i K(x_i, x) + b \tag{5}$$

with

$$K(x_i, x_j) \equiv \varphi(x_i)^T \varphi(x_j) \quad (6)$$

being the kernel function, $a_i$ the weighting factors and b the bias term. In our case the kernel is a radial basis function (RBF), which is defined as

$$K(x_i,x_j)=\exp(-\gamma\|x_i-x_j\|^2) \geq 0 \quad (7)$$

where

$$\gamma = \frac{1}{2\sigma^2} \quad (8)$$

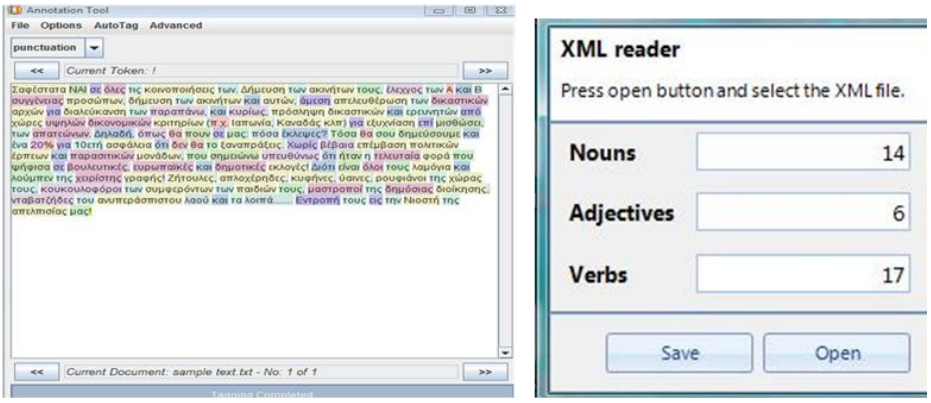
($\sigma$ is the standard deviation) is a parameter on the kernel.

The RBF kernel non-linearly maps samples into a higher dimensional space, so it can handle cases when the relation between class labels and attributes is non- linear. The parameters $\gamma$ and C were defined heuristically. In our application we have used the SVM training algorithm provided by the LIBSVM library [1].

In order to increase our classification results, a Correlation Feature Selection (CFS) procedure is used to rank the extracted features. The CFS algorithm, proposed by Hall [5] is based in the central hypothesis that good feature sets contain features that are highly correlated with the class (valid, not valid), yet uncorrelated with each other. CFS is a filter approach independent of the classification algorithm by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.

## 4 Evaluation and Experiments

To evaluate the proposed methodology data derived from 297 users' comments (posts) published on the Nafteboriki newspaper's blog (http://www.naftemporiki.gr/debates/) were collected. The comments were written about a certain subject about "the issue of publishing the names of people who don't pay their taxes or not", and they concern a two months time period, during which the Greek government would decide if the decision would be implemented. A comment can be added by any user, anonymously or not, even when he is not a subscriber for the newspaper. An experienced sociologist after reading carefully all posts, annotated each one of the as "high opinion strength" if the user support his/her opinion using arguments or "low opinion strength" otherwise. The expert's opinion is used as a golden standard for our classification schema. A freely available tagger software initially created by ¨Natural Language Processing Group Department of Informatics - Athens University of Economics and Business (http://nlp.cs.aueb.gr/software.html), is used to characterize every part of each post as noun, adjective, verb or punctuation symbol. This software automatically tags nouns, adjectives, articles, verbs, conjunctions and adverbs using different colors as shown in Figure. 1a




(a) (b)

**Fig. 1.** (a) The Tagger software automatically tags nouns, adjectives, articles, verbs, conjunctions and adverbs using different colors. (b) A script is used to count the total number of nouns, adjectives, verbs and punctuation symbols per post

In order to count (for every post) the total number of nouns, adjectives, verbs and punctuation symbols a script is prepared (Figure. 1b). Finally, an automated word count software is used to count the number of words in each post. Using the above described procedure the following features are extracted for each post (Table 1):

**Table 1.** Extracted features

| Feature # | Feature description |
|---|---|
| 1. | # of words per comment. |
| 2 | # of nouns divided by the # of words per comment. |
| 3 | # of adjectives divided by the # of the words per comment. |
| 4 | # of verbs divided with the # of the words per comment. |
| 5 | The spelling mistakes divided with the # of the words per comment. |
| 6 | Usage of uppercase letters or not (usage designated as 1, where lack of usage was designated as 0). |
| 7 | Usage of punctuation symbols i.e. dots, commas, interrogation marks etc (usage designated as 1, where lack of usage was designated as 0). |

The constructed dataset consists of the above mentioned features extracted for 297 posts. 186 of them are classified by the expert as “high opinion strength” and 111 are classified as “low opinion strength”.

Having applied the RBF kernel SVM algorithm in our dataset the initial classification accuracy is 73.06%. The confusion matrix of the classification problem is provided in Table 2.

**Table 2.** Confusion matrix produced using all available features

| Class | Classified as low opinion strength | Classified as high opinion strength |
|---|---|---|
| Low opinion strength | 62 | 49 |
| High opinion strength | 31 | 155 |

To enhance our classification results we have applied the CFS feature selection algorithm. The best ranked features are shown in Table 3.

**Table 3.** Confusion Best ranked features according to CFS algorithm

| Ranking # | feature |
|---|---|
| 1 | # of words per comment. |
| 2 | # of nouns divided by the # of words per comment. |
| 3 | # of verbs divided with the # of the words per comment. |
| 4 | Usage of uppercase letters or not |

After selecting only the top ranked features we apply again the SVM classifier. The obtained accuracy is 78.11%. Table 4 provides the new classification problem confusion matrix. Finally Figure 2 provides a graphical representation of the two classification schemas comparative results in terms of accuracy.

**Table 4.** Confusion matrix produced using CFS selected features

| Class | Classified as low opinion strength | Classified as high opinion strength |
|---|---|---|
| Low opinion strength | 65 | 46 |
| High opinion strength | 19 | 167 |

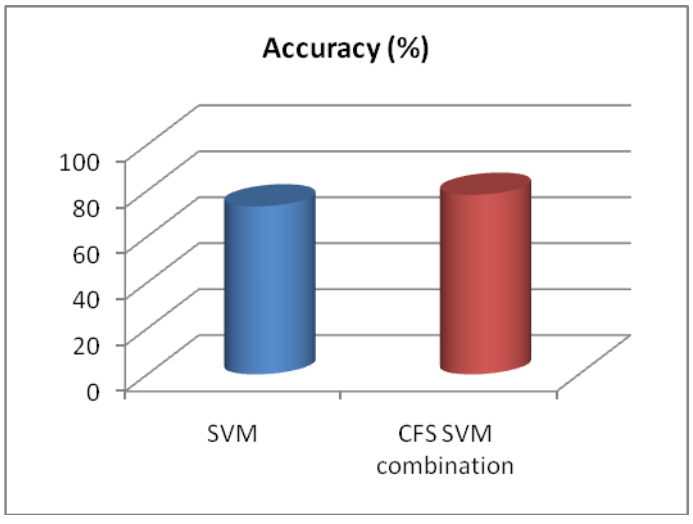


**Fig. 2.** Classification comparative results in terms of accuracy.

## 5 Concluding Remarks

We have presented an innovative methodology that is able to automatically extract quantitative features from text web sources (e.g blogs) and classify the user's opinions strength as "high" if the user supports his opinion using strong arguments or "low" otherwise. To validate the proposed methodology we have constructed a database consisting of features extracted of 297 posts arising from a Greek newspaper blog. The initial experimental results are promising.

Our future works includes the enrichment of our dataset, the employment of more advanced classifiers in order to increase our classification accuracy and the testing of our methodology into real world posts dealing with different topics.

In addition, the future work will aim at implementing fully integrated software which could deliver efficient opinion mining automatically using the SVM algorithm and the tagger features in a homogenous software environment. This tool should allow the users to implement effective opinion mining tasks given a blog, forum or a social network via a usable web interface. The tool, its capabilities, features, functional and technical specifications are in an early development phase.

## References


1. Chih-Chung Chang and Chih-Jen Lin, 2001, LIBSVM : a library for support vector machines.
2. Decman. Mitja, 2009. "Web 2.0 in e-Government: The challenges and opportunities of Wiki in Legal Matters". Proceedings of the 9th European Conference on e-Government, pp 229-236.
3. Furuse O., Hiroshima N. Yamada S.and Kataoka R., 2007, "Opinion sentence search engine on open-domain blog", Proceedings of the 20th international joint conference on Artifical intelligence in Hyderabad, India, pp.2760-2765
4. Gribble D. S.et al., 2000 "Scalable, Distributed Data Structures for Internet Service Construction," Proc. 4th Symp. Operating Systems Design and Implementation, Usenix Assoc., pp. 319–332.
5. Hall M.A., 1998, "Correlation-based Feature Subset Selection for Machine Learning" Hamilton, New Zealand, 1998.
6. Hatzivassiloglou, V., and Wiebe, J. 2000. "Effects of adjective orientation and gradability on sentence subjectivity." In proceedings of the Conference on Computational Linguistics.
7. Hayat. Ali, Linda Macaulay and Liping Zhao. 2009. "A Collaboration Pattern Language for e-Participation":A Strategy for Reuse". Proceedings of the 9th European Conference on e-Government, pp 29-39.
8. Jian-Tao Sun, Xuanhui Wang, Dou Shen, Hua-Jun Zeng, and Zheng Chen. 2006 "Cws: A comparative web search system." In International Conference on World Wide Web (WWW), 2006

9. Ku LW Chen H.H 2007 “Mining opinions from the Web: Beyond relevance retrieval Source”, Journal of the American Society for Information Science and Technology Volume 58, Issue 12,October 2007
10 Pang Bo and Lee Lillian, 2008. “Opinion Mining and Sentiment Analysis.” Foundations and Trends in information Retrieval. Vol 2 Nos. 1-2, pp1-135. DOI:10.1561/1500000001.
11 Pang, B.; lee, L.; and Vaithyanathan, S. 2002. “Thumps up? Sentiment classification using machine learning techniques”. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-02)
12 Qingliang Miao, Qiudan Li and Ruwei Dai, 2009 “AMAZING: A sentiment mining and retrieval system”, Expert Systems with Applications
13 Turney, P. 2002. “Thumps up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the association for Computational linguistics.
14 V. Vapnik,1995, “The nature of statistical learning theory”, Springer, New York.
15 Wiebe, J., Wilson, T.; Bruce, R.; Bell, M.; and Martin, M 2002. “Learning subjective language.” Technical Report TR – 02-100, Department of Computer Science, university of Pittsburgh, Pittsburgh, Pennsylvania.
16 Wiebe, J.; Bruce, R.; and O’Hara, T. 1999. Development and use f a gold standard data set for subjectivy classifications. In Proceedings of the 37th Annual Meeting of the Association for Computational Lingyistics (ACL-99), 246-253.
17 Wiebe, J.2000. “Learning subjective adjectives from corpora”. In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI -2000).
18 Yu, H., and Hatzivassiloglou, V 2003. “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences”. In Proceedings of the Conference on empirical Methods in Natural language Processing
19 Z. Xu and R. Ramnath, 2009 “Mining Opinion from Poll Results in Web Pages,” WWW2009, April 20-24, 2009, Madrid Spain
20 Zabin J and Jefferies A, 2008 “Social media monitoring and analysis: Generating consumer insights from online conversation”, Aberdeen group Benchmark Report.
21 Zhongwu Zhai, Bing Liu, Hua Xu and Peifa Jia, 2011 "Clustering Product Features for Opinion Mining." to appear in Proceedings of Fourth ACM International Conference on Web Search and Data Minin, Hong Kong, China.