Web-based experimental platform for sentiment analysis

Jasmina Smailović[1], Martin Žnidaršič[2], Miha Grčar[3]

ABSTRACT

An experimental platform is presented in the paper, which is used for the evaluation of various approaches of automatic sentiment analysis in financial texts. The platform provides an experimental environment which enables the user to quickly assess the behavior of various algorithms for sentiment detection in given text from an arbitrary HTML source. There is a basic sentiment word tagger and country tagger available to facilitate the preview of text under analysis. Development of the platform and the algorithms it hosts is still in progress and is intended for experimental purposes only. However, as it is implemented as a publicly available web application, it can find its use also in general public with an interest in sentiment revealing parts of HTML sources.

Key Words: *sentiment detection, semantic enrichment, textual analysis, word tagging, classification, web-based application.*

1 INTRODUCTION

Individuals, companies, political parties, movements have always been interested in how the world sees their actions, their brands, their products, them. Until a decade or two ago, the only approximation of the "world" were the media and the only way to get an insight into its viewpoint was manual clipping, laborious manual gathering of news related to a person, action, company, brand or any other such entity. With the recent transfer of news from paper to the Web, the tapping into world's viewpoint became easier. Even more, with massive amounts of user–generated contents, such as blogs, forums and social networks, the sources of the insights into the viewpoints of the “world” became much more abundant, accessible and relevant.

Therefore, it is not unusual that development of methodologies, solutions and products for automatic gathering and analysis of such information is currently in its full extent. In this paper, we present a Web-based platform that provides an experimental environment for testing various sentiment detection algorithms. The platform enables use and assessment of predefined sentiment classifiers on arbitrary HTML sources. To facilitate the preview of text under analysis, there is also a sentiment word tagger and country tagger. The platform is meant to serve for experimental comparison of various approaches of sentiment detection and analysis. However, as it is implemented as publicly available web application, it might find also some other use by the interested public.

The paper is structured as follows: sentiment analysis is briefly introduced in section 2, followed by a presentation of our solution in section 3. Conclusions and directions for further work are given in the final section 4.

[1] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, corresponding author, jasmina.smailovic@ijs.s*

[2] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, martin.znidarsic@ijs.s*

[3] *Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, miha.grcar@ijs.s*

2 SENTIMENT ANALYSIS

Other people`s opinion has always been an important piece of information during decision-making processes. The Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of a vast pool of people that are neither someone's personal acquaintances nor well-known professional critics - that is, complete strangers. And conversely, more and more people are making their opinions available to the public via the Internet (McCallum and Nigam, 1998). With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, social networks and various other types of social media, all of which continue to grow at lightning speed, consumers now have at their disposal a highly effective soapbox by which to share their brand experiences and opinions, positive or negative, regarding any product, brand or service (Zabin and Jefferies, 2008).

Sentiment analysis or opinion mining refers to the application of natural language processing[4], computational linguistics[5], and text analytics[6] to identify and extract subjective information in source materials. The term "sentiment" used in reference to the automatic analysis and tracking of the predictive judgments therein first appears in 2001 papers by Das and Chen (2001) and Tong (2001), due to these authors' interest in analyzing market sentiment. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. The attitude may be his or her judgment or evaluation, affective state (i.e. the emotional state of the author when writing), or the intended emotional communication (i.e. the emotional effect the author wishes to have on the reader).

A classic sentiment analysis application is summarization and analysis of customers reviews; for example, tracking what bloggers are saying about a brand like Toyota or Google. Sentiment analysis can also be used for enabling technologies for other systems; detection of "flames" (overly heated or antagonistic language) in email or other types of communication; in online systems that display ads as sidebars. Sentiment analysis can be used to detect webpages that contain sensitive content inappropriate for ads placement; or it can be used as an augmentation to recommendation systems, since it might behoove such a system not to recommend items that receive a lot of negative feedback (McCallum and Nigam, 1998). Next, sentiment analysis can be used in business intelligence; according to consultant Tom H.C. Anderson of Anderson Analytics, the analysis process applied in studying a brand of Dove company starts with surveys and web scraping from online consumer forums. Such sites related to the brand and the brand campaigns have many thousands of messages with potential value. Anderson's company codes and characterizes data, looking for sentiment polarity – positive, negative and neutral – seeking to understand emotions and attitudes (Grimes, 2008).

Automatic sentiment detection and analysis can be conducted by employing expert-based natural language models or machine learned models to the documents of interest. In simpler approaches, the documents are checked for specific words, phrases, word relations and other similar features of text. These might be simply reported or used as a basis for designating document's sentiment type. The so−called deeper approaches try to understand the written text

[4] Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages.

[5] Computational linguistics is an interdisciplinary field dealing with the statistical and/or rule-based modeling of natural language from a computational perspective.

[6] The term text analytics describes a set of linguistic, statistical and machine learning techniques that model and structure the information content of textual sources for business intelligence, exploratory data analysis, research, or investigation.

and its implications through reasoning tools and ontologies. The other automated approach is sentiment detection with machine learning methods. These require a labeled dataset, i.e. a collection of documents with prescribed sentiment type (target class). Machine learning techniques use such a dataset to construct a model that allows for classification of unlabeled documents into one of sentiment types, called target classes in general. The classification depends on the document's features, which are usually frequencies of words and word combinations in the document. The features, however, can be arbitrarily complex and it has been shown many times that the feature creation part is a crucial step in text classification procedure.

3 WEB-BASED PLATFORM FOR SENTIMENT ANALYSIS

Our approach towards sentiment analysis, a platform for experimenting with word tagging and sentiment detection, is described in the following two subsections. The platform consists of an input component that prepares text from an arbitrary URL for analysis and two sentiment analysis tools: a simple sentiment word tagger and a collection of sentiment detectors or classifiers. The sentiment word tagger is described in Section 3.1, while the experimental sentiment detection set is described in Section 3.2.

Our application is web-based, therefore it runs in any standards compliant web browser. It is available from the following URL:

http://first.ijs.si/demos/sentag/

The application is implemented in ASP.NET Framework 4 in C# on top of LATINO[7], a software library that provides a range of data mining and machine learning algorithms with the emphasis on text mining, link analysis, and data visualization. For some parts of the application we used JavaScript Library, i.e. jQuery; for example, for sentiment word and country tagging.

3.1 Sentiment word and country tagger

Tagging is a process of marking up the words, process or other components in text. It is useful for facilitating the preview of text under analysis and can serve also the purpose of further automatic text processing. We have created an application that tags single positive and negative sentiment bearing words in text from a given URL source. The words in text are tagged for sentiment simply according to their appearance in positive and negative word lists described in the paper by Loughran and McDonald (2011), which are also available online[8]. There is also a country tagger in our application, which tags country names on the basis of their appearance in the UN list of countries and areas[9]. In the current version we did not take into account usage of negators such as *not* and *never*, which flip the polarity of a word, or modal operators as *might*, *could* and *should*, which distinguish hypothetical from real situations.

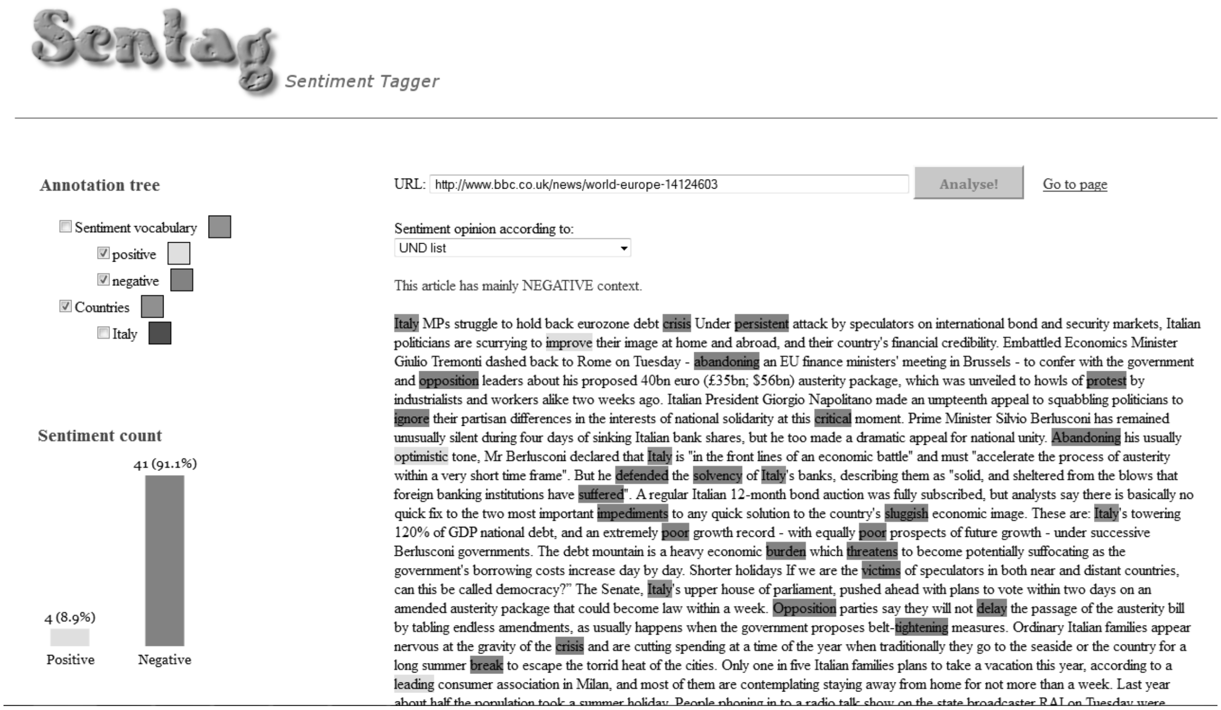
In Figure 1, an example of tagged text is given. As can be seen from the figure, on the top of the web page there is text field where the user can type the URL of a desired web page containing text that is to be analyzed. The user needs to click on "Analyse!" button to display the text of the article. On the left side of the page an annotation tree is shown. There is sentiment

[7] LATINO (Link Analysis and Text Mining Toolbox) is open-source—mostly under the LGPL license—and is available at http://latino.sourceforge.net/

[8] http://www.nd.edu/~mcdonald/Word_Lists.html

[9] http://unstats.un.org/unsd/methods/m49/m49alpha.htm

vocabulary with checkboxes for positive and negative sentiments in the upper part of the tree. The user can use the checkboxes to highlight positive, negative, or both sentiments in the text. Positive sentiments are always highlighted with yellow color and negative ones with red color. In the lower part of the tree there is a list of countries that are mentioned in the article. The user can highlight countries by checking the checkbox on the left side of a country`s name. Colors for countries are random. There is also a possibility to highlight countries all at once with the same color by checking the checkbox next to the node “countries” in the annotation tree. Additionally, below the annotation tree, there is a chart that shows how many positive and negative sentiment bearing words are found in the text.



URL: http://www.bbc.co.uk/news/world-europe-14124603

Sentiment opinion according to: UND list

This article has mainly NEGATIVE context.

Italy MPs struggle to hold back eurozone debt crisis Under persistent attack by speculators on international bond and security markets, Italian politicians are scurrying to improve their image at home and abroad, and their country's financial credibility. Embattled Economics Minister Giulio Tremonti dashed back to Rome on Tuesday - abandoning an EU finance ministers' meeting in Brussels - to confer with the government and opposition leaders about his proposed 40bn euro (£35bn; $56bn) austerity package, which was unveiled to howls of protest by industrialists and workers alike two weeks ago. Italian President Giorgio Napolitano made an umpteenth appeal to squabbling politicians to ignore their partisan differences in the interests of national solidarity at this critical moment. Prime Minister Silvio Berlusconi has remained unusually silent during four days of sinking Italian bank shares, but he too made a dramatic appeal for national unity. Abandoning his usually optimistic tone, Mr Berlusconi declared that Italy is "in the front lines of an economic battle" and must "accelerate the process of austerity within a very short time frame". But he defended the solvency of Italy's banks, describing them as "solid, and sheltered from the blows that foreign banking institutions have suffered". A regular Italian 12-month bond auction was fully subscribed, but analysts say there is basically no quick fix to the two most important impediments to any quick solution to the country's sluggish economic image. These are: Italy's towering 120% of GDP national debt, and an extremely poor growth record - with equally poor prospects of future growth - under successive Berlusconi governments. The debt mountain is a heavy economic burden which threatens to become potentially suffocating as the government's borrowing costs increase day by day. Shorter holidays If we are the victims of speculators in both near and distant countries, can this be called democracy?” The Senate, Italy's upper house of parliament, pushed ahead with plans to vote within two days on an amended austerity package that could become law within a week. Opposition parties say they will not delay the passage of the austerity bill by tabling endless amendments, as usually happens when the government proposes belt-tightening measures. Ordinary Italian families appear nervous at the gravity of the crisis and are cutting spending at a time of the year when traditionally they go to the seaside or the country for a long summer break to escape the torrid heat of the cities. Only one in five Italian families plans to take a vacation this year, according to a leading consumer association in Milan, and most of them are contemplating staying away from home for not more than a week. Last year about half the population took a summer holiday. People phoning in to a radio talk show on the state broadcaster RAI on Tuesday were

Figure 1: Positive and negative sentiment bearing words and countries tagged in text of an article. The article is taken from BBC News web site, http://www.bbc.co.uk/news/world-europe-14124603

3.2 Sentiment detection/classification

In written text (Toivanen, Vayrynen and Seppanen, 2004), the smallest unit of sentiment detection is a single word. Bigger units can be phrases, sentences, paragraphs, documents and document collections. Detection of sentiment on a word-basis can be relatively straightforward, provided that we have lists of sentiment bearing words for a particular domain and some basic text analysis tools, such as tokenizers, lemmatizers, etc. The issue of the word-lists' target domain is important, since it has been shown (Loughran and McDonald, 2011) that general purpose sentiment word-lists are of very limited use in some particular domains. There are further challenges for sentiment detection, however, such as the issue of ambiguous words with multiple (sometimes opposed) meaning, or the issue of writing styles such as sarcasm that reverse the words' meanings. Even single-word sentiment detection becomes a challenging task in such situations, demanding non-trivial analysis of word's context for a proper sentiment designation.

Sentiment detection is an important step in sentiment analysis. It can be the main goal of the analysis itself in cases when insights into the sentiment's context, sources and other relations that would demand further analysis are not necessary.

Our application provides a general designation of the document’s sentiment. This is achieved with classifiers which are based on word-list or on models learned from examples of datasets with known positive or negative sentiment. Currently, our application supports sentiment detection with three approaches: (1) word-list-based sentiment detection, (2) sentiment detection using Naïve Bayes classifier and (3) sentiment detection using KNN classifier.

Word-based sentiment detection provides a general designation of the document’s sentiment into one of the three possible classes: *positive*, *negative* or *complicated*. The document is classified as positive, if over 70% of sentiment bearing words were found to be of positive sentiment. Similarly, if more than 70% of such words were found to be of negative sentiment, the document will be classified as negative. In all other cases, the document is classified as complicated, which means that it is difficult to determine the sentiment of the document. In Figure 1, an example of a document classified as negative is shown.

In machine learning and pattern recognition, classification refers to a procedure for assigning a given piece of input data into one of a given number of categories. An algorithm that implements classification is known as a classifier. The piece of input data is formally called an instance, and the categories are called classes. The instance is described by a vector of features, also called attributes, which together constitute a description of all known characteristics of the instance. Classification normally refers to a supervised procedure, i.e., a procedure that learns to classify new instances based on learning from a training set of instances that have been properly labeled by hand with the correct classes. The corresponding unsupervised procedure is known as clustering, and involves grouping data into classes based on some distance/similarity metric.

Naïve Bayes classifier in our application classifies text in one of the two possible classes: *positive* or *negative*. The Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naïve) assumption of attributes’ independence. While this assumption is clearly false in most real-world tasks, naïve Bayes classifier often performs very well (McCallum and Nigam, 1998). Because of independence assumption, the parameters for each attribute can be learned separately, and this greatly simplifies learning, especially when the number of attributes is large, e.g. like in our case of document classification where the attributes are words and word pairs, whose numbers are large.

Similarly as the Naïve Bayes classifier, KNN classifier in our application also classifies text in one of the two classes: *positive* or *negative*. The KNN classifier is an instance-based classifier which operates on the assumption that classification of unknown instances can be done by relating the unknown to the known according to some distance/similarity metric. The intuition is that two instances far apart in the instance space defined by the appropriate distance function are less likely to belong to the same class than two closely situated instances.

By using different datasets for learning the classifiers, the results obtained are different, although the learning algorithm stays the same. For training Naïve Bayes and KNN classifier we used a dataset of 375,856 sentiment labeled tweets which is described in the paper by Saveski and Grčar (2011). In the future, we plan to use our platform to experiment with various classifiers and datasets.

After choosing a desired approach for classification from the drop down list and clicking the “Analyse!” button, above the text of the article one additional sentence is displayed which shows if the article has mostly positive or negative context or it is complicated to determine context. The font color is red for negative context, green for positive context and gray for context which is complicated.

4 CONCLUSIONS

In this paper, we have described our approach towards experimenting with sentiment detection and tagging. We have created a web-based application that provides an experimental environment for testing of various sentiment detection algorithms. The application serves for experimental comparison of various approaches of sentiment analysis. Also, it can find its use in general public because it is publicly available.

Development of the application is still in progress and there are numerous possible improvements, from simple to very complex. To start with simpler ones, we first plan to add company names tagger, similarly as it is currently for the countries. Next, negation detection functionality will have to be provided, which can be done in a number of different ways. Then, we plan to experiment with different datasets for training classifiers and to evaluate accuracy. The final, the most difficult and the most interesting problem that we plan to tackle on the presented platform is sarcasm detection. We expect to experiment with a number of natural language analysis and machine learning tools for this purpose and with combinations of both.

ACKNOWLEDGMENTS

The work presented in this paper has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) within the context of the Project FIRST, Large scale information extraction and integration infrastructure for supporting financial decision making, under grant agreement n. 257928 and by the Slovenian Research Agency through the research program Knowledge Technologies under grant P2-0103.

REFERENCES

Das, S., Chen, M. 2001. *Yahoo! for Amazon: Extracting market sentiment from stock message boards*, in Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Grimes, S. 2008. *Sentiment Analysis: A Focus on Applications*, BeyeNETWORK - Global coverage of the business Intelligence Ecosystem.

Loughran, T., McDonald, B., 2011. *When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks*, Journal of Finance Journal of Finance, Vol. 66, pp 35-65.

McCallum, A., Nigam, K. 1998. *A comparison of event models for Naive Bayes text classification*, Computer and Information Science, Vol. 752, pp. 41-48

Pang, B., Lee, L. 2008. *Opinion Mining and Sentiment Analysis*, Foundations and Trends(R) in Information Retrieval, Vol. 2, Nos. 1–2, pp 1–135.

Saveski, M., Grčar, M. 2011. *Web Services for Stream Mining: A Stream-Based Active Learning Use Case*, to appear in proceedings of PlanSoKD ECML-PKDD Workshop.

Toivanen, J., Vayrynen E., Seppanen, T. 2004. *Automatic Discrimination of Emotion from Spoken Finnish*, Language and Speech, December 2004 vol. 47 no. 4 383-412.

Tong, R. M. 2001. *An operational system for detecting and tracking opinions in on-line discussion*, Proceedings of the Workshop on Operational Text Classification (OTC).

Zabin, J., Jefferies A. 2008. *Social media monitoring and analysis: Generating consumer insights from online conversation*, Aberdeen Group Benchmark Report.