

# Sentiment Analysis through Machine Learning: an Experimental Evaluation for Albanian

Marenglen Biba<sup>1</sup> and Mersida Mane<sup>2</sup>

**Abstract.** Opinions have always influenced our behaviours and they have a key role in human activities. Nowadays, online opinion resources such as newspapers, blogs, and reviews have enormously increased the amount of text data available for analysis. Sentiment Analysis (or Opinion Mining) is increasingly becoming an important tool for analysing text data in order to understand opinions correctly. In this context, machine learning methods have the potential to perform correct classification of texts as expressing positive or negative opinion for a certain topic. However, much research has been dedicated to languages such as English, Japanese, Chinese or German but no research has been made for other rare Indo-European languages such as the Albanian. In this paper, we present the first approach for Sentiment Analysis in Albanian. We show through extensive experiments with text data from political news consisting of five different topics, that the proposed approach is effective in classifying text documents as belonging to negative or positive opinion regarding the given topic.

## 1 Introduction

What other people think is increasingly becoming important for many decision-making processes [8]. Opinions have always influenced our behaviour and they have a key role in human activities. Long before, many of us looked for opinions by asking our friends and family. Consumers always asked other users of a product or service before doing a purchase. On the

---

<sup>1</sup> Marenglen Biba

University of New York in Tirana, Albania, email: marenglenbiba@unyt.edu.al

<sup>2</sup> Mersida Mane

University of New York in Tirana and University of Greenwich, London, UK  
email: mersidamane@unyt.edu.al

other hand, whenever an organization had to make decisions based on the opinions of the consumers about its products or services, it conducted surveys, focus groups or opinion polls.

With the explosive growth in the past few years of the social media content (e.g., blogs, forum discussions and posting in social network sites) on the Web, worldwide communication has changed [6]. We are not any more limited to ask our friends or family for their opinions. There are a great number of opinion-rich resources, such as user reviews on Web, forums, blogs and social sites, where individuals that are neither our acquaintances nor professional critics discuss with each other and express their opinions. There is no more need for the organizations to gather customer opinions by conducting surveys, focus groups or opinion polls, since there is plenty of information publicly available on the Web.

Sentiment analysis is an approach for extracting feelings from a given text. For this task to be effectively accomplished, it is required to process huge amounts of data available online and to retrieve the opinion hidden in it. However, in processing subjective information effectively by these systems, there are a number of challenges to overcome [8]:

First, applications integrated into a general-purpose search engine require determining if the user is looking for subjective material. This could not be a difficult problem if the queries contain indicator terms like “opinion”, “review”, or the application provides a checkbox to the users giving them the possibility to indicate that reviews are desired. However, query classification remains still a difficult problem.

Second, besides the still-open problem of determining which documents are topically relevant to an opinion-oriented query, an additional challenge is determining which documents or portions of documents contain review-like or opinionated material. For texts fetched from review aggregations sites like ‘Amazon.com’ or ‘Epinions.com’, this is a relatively easy problem because the review-oriented information has a stereotyped format. However, the material in other sites can vary widely in presentation, style, content and grammatical level.

Third, it is hard for computers to analyze free-form texts because of additional challenges such as quotations included in an article. The views expressed in each quotation must be attributed with the correct entity.

Finally, the sentiment information will be presented in a reasonably summarized fashion, for example, by highlighting some of the opinions, through representation of points of consensus and disagreement or by considering the level of authority of the opinion holders. Sometimes producing a visualization of sentiment data is more appropriate than a textual summary of it.

The last three challenges represent also the most active areas of research. In addition, these challenges highlight some concerns that are also faced during sentiment analysis in Albanian.

Significant research work on sentiment analysis has been done on languages such as English, Japanese, Chinese, German, and Romanian [2]. So far, to the best of the authors' knowledge, there have been no sentiment analysis approaches for Albanian. In this paper we build the first dataset tagged with sentiment information, and perform extensive experiments with several algorithms. We use in our approach a stemmer for Albanian in order to bring the words and their variations to the original form which is the stem or root. This greatly helps in reducing the number of terms that are used in the next stage for text classification. We show that by providing a sound and significant dataset for training, it is possible to achieve high results in the classification of texts expressing a certain opinion into the right category as being positive or negative.

The paper is organized as follows: Section 2 introduces Sentiment Analysis and related approaches. Section 3 contains a brief introduction to Albanian with some features that make it a complex language to deal with, Section 4 contains the experimental results and we conclude in Section 5.

## 2 Sentiment Analysis

In this section we introduce Sentiment Analysis and related issues.

### 2.1 The Problem of Sentiment Analysis

Everyone can express his own opinion about everything that might be for example a service, a product, an event or a topic, which is called target object or entity. The authors of an expressed opinion are known as opinion holders or opinion sources. Opinions fall into one of these two groups: regular opinions and comparative opinions. Regular opinions are known as opinions in the research literature [1]. Comparative opinions express preferences of the opinion holder in relation to two or more target objects based on their aspects. An opinion orientation can be positive, negative or neutral which are also known as polarities or semantic orientations [5].

A regular opinion is a *quintuple*  $(e_i, a_{ij}, oo_{ijkl}, h_k, t_l)$ , where  $e_i$  is the name of an entity,  $a_{ij}$  is an aspect of  $e_i$ ,  $oo_{ijkl}$  is the orientation of the opinion about aspect  $a_{ij}$  of entity  $e_i$ ,  $h_k$  is the opinion holder, and  $t_l$  is the time when the opinion is expressed by  $h_k$  [1]. All the above components are important because if any of them is missing, it becomes complicated in general. For

example, in the sentence ‘The touch screen was perfect’, it is impossible to decide to whom this touch screen belongs to, so this opinion is worthless. However, there are cases when not all the components are required. For example, when required to know the opinions of thousands of people, the opinion holders component is not needed. Also, there are cases when new components are needed to be added to the quintuple, such as the gender and the age of the opinion holders. This kind of definition for opinion plays a key role for the transformation of unstructured text to structured data. It helps with important information for both qualitative and quantitative analysis of opinions.

Two key concepts in opinion mining are subjectivity and emotion. As mentioned in [6], an objective sentence consists of facts, while a subjective sentence consists of personal feelings, attitudes or beliefs. Subjective expressions might be opinions, wishes, viewpoint, doubts, allegations or speculations and very often they do not offer opinions [10, 14]. Also, not all the objective sentences offer an opinion. For example, the objective sentence “*the battery of my laptop lasts 45 minutes*” does mean a negative opinion. The association of subjectivity with opinion is still confusing for the researchers. Even though subjective sentences are different from opinion sentences, there is great intersection between them.

Emotions are spontaneous feelings and thoughts. The six main emotions are love, happiness, surprise, anger, sorrow and fright which include a number of other secondary and tertiary ones [9]. The intensity of the emotion defines how strong an opinion is. There are a lot of cases when an opinion sentence has no emotion in it and when an emotion sentence contains no opinion.

## 2.2 Aspect-Based Opinion Summary

The number of opinions used for the majority of opinion mining application is significantly high. Opinions from one opinion holder are usually not enough. This means that in such cases it is preferable a summary of opinions. Opinion quintuples offer very useful resources of information for obtaining qualitative and quantitative summaries. Aspect-based opinion summary is an ordinary type of summary based on aspects [1].

Aspect-based opinion summary can be presented by using bar charts or text summary. Text summary is a brief overview of what the others think related to a product or service. The main disadvantage of text summary is that it is only qualitative where in cases such as analytical purposes it is not helpful. For example, a common text summary would be “Lots of people think that Apple computers are the best”, while the quantitative sum-

mary for the same sentence would be “87% of people think that Apple computers are the best, while 13% prefer other brands”. Quantitative presentation is as fundamental to many opinion mining applications as the traditional survey researches. In this context a lot of research is dedicated to changing the text summary into a more readable form.

### ***2.3 Document-Level Sentiment Classification***

In a document-level sentiment classification the whole document is considered as the fundamental information component for performing the classification and deciding if it expresses a positive opinion, negative opinion or sentiment [1].

Sentiment classification considers the document as expressing opinions for only one entity and the opinions are expressed from only one opinion holder. This is true for customer reviews because they are only for a product or service and are expressed by only one reviewer. In forums or blog postings this is not true because one can express his opinion for many products or services and can also compare these products or services by using comparative sentences.

### ***2.4 Supervised and Unsupervised Learning for Sentiment Analysis***

Sentiment classification can be described as a supervised learning problem which consists of these three classes: positive, negative and neutral. Product or service reviews are the common data used for training and testing. These rated reviews, for example from 1 to 5 stars, serve as ready data for further training and testing. For example, reviews assigned 1-2 stars are thought to be negative reviews, reviews assigned 3 stars are thought to be neutral ones and those reviews assigned 4-5 stars are thought to be positive reviews. Any of the current supervised learning methods can be used for sentiment classification.

The authors in [8] have shown that the usage of features such as unigrams in classification had the same performance by using both methods Naive Bayes and Support Vector Machines. Some of the most important features are: Terms and their frequency, Part of speech, Opinion words and phrases, Negations, Syntactic dependency. Different types of approaches are used to increase the classification accuracy, for example the use of a score function [4], feature weighting schemes [6], or the utilization of opinion words during the training procedure [12].

Opinion words and phrases are the key indicators for sentiment analysis, thus applying unsupervised learning methods based on these words and

phrases is normal. An unsupervised approach is presented in [14], which has three main steps: in the first step adverbs and adjectives are extracted since they are helpful containers of opinions; in the second step, it is computed an equation for calculating the pointwise mutual information; in step three, for a given text, the algorithm calculates the average semantic orientation (SO) for all the phrases and based on the value of the average SO, the text is classified.

### 3. Albanian Language

Nowadays, Albanian is a language spoken in Albania, as well as in Kosovo, northwest of Macedonia, southwest of Montenegro and northwest of Greece. It is spoken even wider, in states such as Italy, Greece, USA and other countries where thousands of Albanians have emigrated and currently live [3]. During the years it has been influenced by other languages due to many invasions. However, Albanian as an Indo-European language has kept its originality with its special structure of phonetic, grammar and lexicon. Albanian alphabet is based on the Latin alphabet and consists of 36 letters, 7 vowels

Albanian is a very rich language in words with more than one meaning which are called polysemantic words. In the context of sentiment analysis these play an important role. For example, the phrase '*pranë zjarrit*' (*near the fire*) can mean really near the fire, but also '*at home*'. In addition it is also a symbol of tranquillity, being safe and a symbol of family life [3]. These cases may mislead an automatic classifier since the opinion being expressed might be related to only one of the meanings of the word.

Lexical field consists of a set of words used to express almost the same idea. For example, the set of words used to express the notion of noises is called lexical field of noises. Words of a lexical field have different types of relations between them. Based on grammatical point of view, in the lexical field of noises the words could be nouns, for example *bubullim* (*thunder*), *fëshfërim* (*whisper*), *zhurmë* (*noise*), and verbs, for example *kërcet*, *thërret* (*call*), etc. Based on semantic point of view, the words could be classified as general, for example *zhurmë* (*noise*), *zëra* (*voices*), and specific words, for example *kuit*, *cinzërit*, etc [3].

Another feature of Albanian is the figurative meaning of words. Many words have several meanings [3]. For example, the word *faqe* (*cheek*) means: the sides of the face (*Lotët i përshkuan faqet* - *Tears ran down her cheeks*); the front part of a hill (*Gjithë faqja e kodrës ishte e mbjellë me ullinj* - *The front of the hill was planted with olive trees.*). However, the

word *faje* can also have the meaning of honor: *si mbeti faje, ska faje ai*. In this last case, the word ‘faje’ is used in figurative meaning.

Homonyms are words that are written or spelled the same, but they have different meanings, for example *mbledhje* which means sum or meeting, depending on the context used [3]. Antonyms are words with opposite meanings. Antonyms are classified as lexical antonyms and grammatical antonyms. Lexical antonyms just have opposite meanings, while grammatical antonyms have similar form, for example: *i ditur (cultured)– i paditur (uncultured)*, *i armatosur (armed) – i çarmatosur (unarmed)*, etc. Antonyms have an important role in expressing opinions.

## 4. Experimental Evaluation

### 4.1 Experimental Setting

Experiments are performed in the Weka framework, a machine learning software written in Java that provides implementation of learning algorithms for different data mining tasks [15]. The workbench includes also many data processing tools. Weka can be used for applying a learning method on a given dataset and evaluate the results, for using learned models to make predictions for new instances and for evaluating the results and comparing the performance of learning algorithms.

Preprocessing tools are used for transforming the instances. They are known as “filters”. Filters are grouped as supervised and unsupervised where supervised filters are used in case class information is available. Both supervised and unsupervised filters are further divided into attribute filters and instance filters.

In the pre-processing step, first, each document is stemmed by using the JStem algorithm [11]. The stemmed documents are then given in input to Weka where each document is transformed into a word vector with the filter `StringToWordVector`.

### 4.2 Corpus Building

For sentiment analysis in Albanian we used political news. They are classified as having positive or negative opinion based on the sentiment of the document. Our corpora consist of five different topics, each containing 40 documents with positive sentiment and 40 documents with negative sentiment. The five topics selected are:

**Topic 1 - Opening of ballot boxes.** An important long and complex debate between the two main parties, Democratic and Socialist in Albania,

has been the opening of ballot boxes for the general elections of 2009. Many news articles can be found on web, where one party claims that the boxes should be opened and the other party insists for the opposite.

**Topic 2 - Immunity reform:** A current topic in Albania is the voting of immunity reform. The Democratic party recommends to the opposite party to vote the immunity reform on August 6<sup>th</sup> 2012, while the opposition leader strongly declares that his party will not vote it until all the proper Constitutional amends are discussed first.

**Topic 3 - Negotiations between Kosovo – Serbia:** Complex relations have been between the Republic of Kosovo and Republic of Serbia, since Kosovo declared its independence in 2008. During 2012 and 2013, several negotiation rounds have been hold between these two countries. However, in Kosovo the opposition is against the negotiations, while the government intends to go ahead.

**Topic 4 - Economic Crisis in Albania:** INSTAT (Institute of Statistics of Albania) and the opposition report about the inflation in different periods during the year expressing negative opinion for the economy. However, there are a few reports which show growth of economy or news where government states that economy is growing. These are used as documents expressing positive opinions.

**Topic 5 - Electoral reform:** No consensus between the two main parties and the smaller parties for voting the electoral reform. The Socialist and the smaller parties have several proposals to achieve the consensus. While the President and the Democratic party are recommending the voting of the electoral reform to be done as soon as possible.

The above topics were selected due to their complexity in terms of the complex language used and also due to the very rich technical vocabulary. In particular, the immunity reform and the electoral reform bear high complexity since in most of the texts of these categories, references to the constitution are made which could probably mislead the classifier. On the other side, most text documents that discuss the economic crisis, contain a high number of quantitative data in terms of percentages, rates etc, which coming in the form of numbers may again mislead the classifier.

### ***4.3 Experimental Results***

In this section we present the experimental results. Table 1 presents all the results obtained with the six different algorithms. We have chosen the most representative algorithms for mining that have been also shown previously to achieve high results in many classification tasks.



Algorithm	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Average
Bayesian Logistic Regression	77.5%	<b>88.75%</b>	86.25%	78.75%	78.75%	83.13%
Logistic Regression	<b>86.25%</b>	87.5%	81.25%	85%	75%	83.00%
SVM	76.25%	87.5%	86.25%	82.5%	82.5%	83.00%
Voted Perceptron	71.25%	82.5%	83.75%	72.5%	72.5%	76.50%
Naïve Bayes	70%	86.25%	<b>88.75%</b>	75%	75%	79.00%
Hyper Pipes	83.75%	87.5%	81.25%	<b>92.5%</b>	<b>90%</b>	87.00%

**Table 1.** Results of mining political news

As we can see from the results by training on the corpus, we are able to produce classifiers with accuracy for the different problems between 86 and 92%. Even though the topics that we have chosen are not easy to deal with due to the complex language and vocabulary used, the approach by first stemming the texts with an Albanian stemmer and then classifying each text with a machine learning algorithm, succeeded in building effective classifiers.

On average, the best performing classifier is HyperPipes. However, depending on the topic, the best performing algorithm for each topic varies. We have used 10-fold-cross-validation in each of the experiments in order to get reliable classification accuracy.

## 5. Conclusion

In this paper we have presented the first attempt for sentiment analysis in Albanian. We have built a corpus to train machine learning models and after stemming every text with an Albanian stemmer, we build classifiers that are able to classify at high accuracy the given texts. Possible future work includes improvements of the stemming algorithm which has a key role in producing the input for the classifier and therefore a direct effect on increasing the value of the performance obtained. In addition, the same experiments could also be performed with a larger dataset and the comparison with the current results would be important to know which should be the minimal or the most appropriate size of the dataset in order to achieve high classification accuracy.

## References

1. Aggarwal, C. C. and Zhai, C. (eds) (2012) *Mining Text Data*, Springer 2012, ISBN 978-1-4614-3222-7.
2. Banea, C., Mihalcea, R. and Wiebe, J. (2011) Multilingual Sentiment and Subjectivity Analysis, In *Multilingual Natural Language Processing*, 2012, Prentice Hall, ISBN 10: 0137151446.
3. Beci, B. (2005) *Gramatika e gjuhes shqipe*. Publisher, Logos-A, 2005. ISBN 9989580197.
4. Dave, K., Lawrence, S. and Pennock, D. (2003) Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proc. of Int'l Conf. on World Wide Web (WWW-2003)*.
5. Jindal, N. and Liu, B. (2006) Mining comparative sentences and relations. In *Proceedings of National Conf. on Artificial Intelligence (AAAI-2006)*.
6. Liu, B. (2010) Sentiment Analysis: A Multi-Faceted Problem IEEE Intelligent Systems, 25(3), 2010, pp. 76-80.
7. Paltoglou, G. and Thelwall, M. (2010) A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2010)*.
8. Pang, B. and Lee, L. (2008) 'Opinion Mining and Sentiment Analysis', *Foundations and Trends in Information Retrieval Journal*, Vol 2.
9. Parrott, W. G. (2001) *Emotions in social psychology: Essential readings*. Psychology Press, ISBN-10: 0863776833.
10. Riloff, E., Patwardhan, S. and Wiebe, J. (2006) Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*.
11. Sadiku, J. (2011). *A Novel Stemming Algorithm for Albanian text in a Data Mining Approach for Document Classification*, Master Thesis. Tirana: University of New York Tirana, 2011.
12. Tan, S., Wang, Y. and Cheng X. (2008) Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-2008)*.
13. Turney, P. D. (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. of the Association for Computational Linguistics (ACL-2002)*.
14. Wiebe, J. (2000) Learning subjective adjectives from corpora. In *Proceedings of Nat'l Conf. on Artificial Intelligence (AAAI-2000)*.
15. Witten, I.H. and Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques* (2<sup>nd</sup> Edition), Morgan Kaufmann Pub, San Francisco.