# HEURISTICS-BASED CLASSIFIER IN A FRAMEWORK FOR SENTIMENT ANALYSIS OF NEWS

**Filip Andonov, Velina Slavova, Marouane Soula**

***Abstract***: *Contemporary systems for sentiment analysis usually work with texts, that are known to be subjective and discovering the sentiment in them is already well studied and developed topic. When dealing with news articles however the main challenge is that the text itself should not be biased, even though that is not granted. Discovering the sentiments that some subjects in the text have about something in it requires more a priori knowledge about the world, the topic, the language, etc. All this requires a system designed for sentiment analysis in financial news to have flexible architecture and knowledge management capabilities. In addition, the social context of the opinion expressed has to be taken into account. That necessitates the establishment of classification criteria concerning the main social characteristics of the opinion holder.*

***Keywords***: *Sentiment analysis, Data mining, Heuristics, Text processing, Gender differences*

***ACM classification keywords***: *I.2 Artificial Intelligence I.2.7: Natural Language Processing, 1.7 Document and text processing*

## Introduction

The goal of the project is to build a data-mining system for sentiment analysis of financial news. Sentiment analysis appeared as an NLP task and the commonly adopted definitions and techniques come from text analysis. A number of extended surveys have been published [e.g., Liu 2010, 2012]. After the overview [Slavova and Hinkov, 2014] we discovered that the problem is in capturing the implicit attitude in objective texts such as the texts of financial news. Most contemporary systems are designed to capture sentiments in the domain of marketing via Internet blogs or other sources in which the opinion holder expresses his opinion openly. Our main problem to overcome is to capture the attitude expressed in the news, despite the fact that news articles are meant to be objective. The problem is how to detect the subtle clues via which the sentiment is transmitted. That is why our strategy includes following the social reaction expressed on the Internet with relation to concrete published news, an event in the domain and the attitude of concrete subjects. This requires a semantic description of the domain, its

actors and events as well as the interconnections and influences between all these ingredients. For that reason our system is based on semantics and tracking of social feedback.

## Short description of the system

The general architecture of the system includes 4 main stages as shown in Figure 1. The articles are retrieved from the Internet and transformed into a convenient format for further treatment. After extraction of several characteristics of the texts by means of a text processing module and natural language processing procedures, each article is assigned to one or more categories. As mentioned, the categorization to an ontological class is necessary in order to discard the articles that are not from the domain, in order to perform the processing related to the domain knowledge.

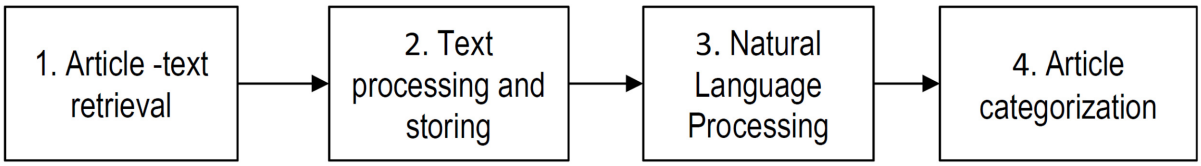


Figure 1. Main architecture of the system

The text processing (Figure 2) is done in the classical way – the text is tokenized in order to separate it into its basic ingredients, i.e. extract the title, paragraphs, sentences and word forms. The content of the article is stored in a database as word-forms (all the strings) and the structure of sentences, paragraphs etc.

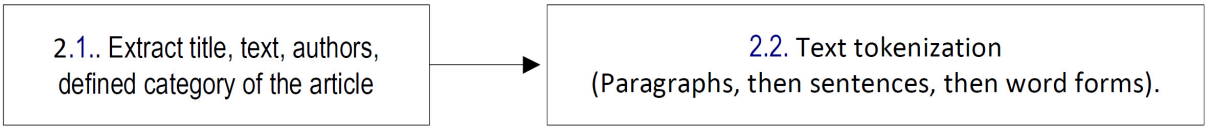


Figure 2. Steps of the text processing module

The Natural Language Processing (NLP) module has two main tasks – to detect the part of speech (POS) for each word form and to identify the Subject of each sentence (figure 3). As the NLP is hard and time consuming task, our strategy is to maintain an Enlarged Dictionary (ED) in the database, which contains the information about the word form itself only once, i.e. what part of speech it is, knowing that a word-form can represent more than one part of speech (table Word-forms in figure 4). In this way the

task of the interconnection between the DB and the texts consists in checking for availability of the word forms in the ED and storage of the extracted characteristics of the sentences and paragraphs only as a structure (word order, order of the sentences and paragraphs). This solution is better from the point of view of memory and processing usage.

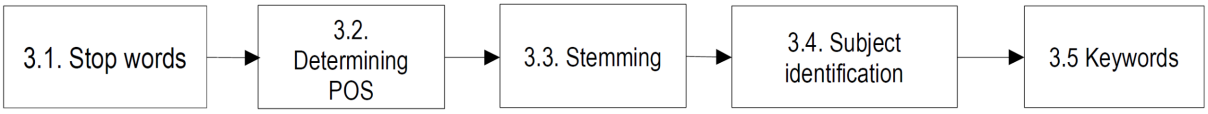


Figure 3. Natural Language Processing module

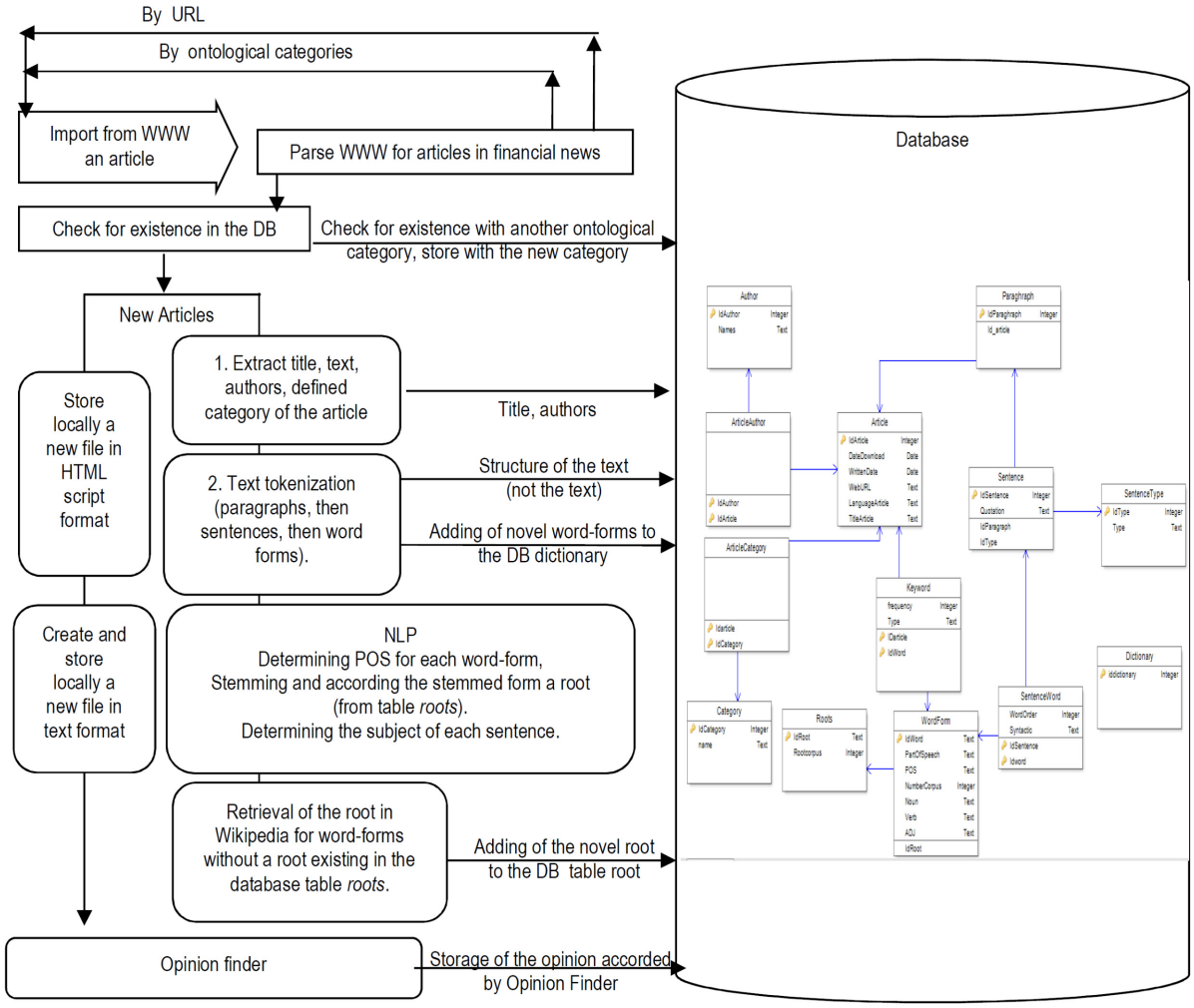


Figure 4. General scheme of the system

The information about Parts of Speech (POS) is necessary for the further sentiment analysis [Nicholls & Song (2009)], namely, as mentioned in the sources, the influence of the different POS on the sentiment detection has to be taken into account.

Concerning the task of Subject identification, it is performed by syntactic tokenization based on the use of Stanford Parser module. We consider the detection of the Subject to be of crucial importance for the further semantic analysis and the detection of the focus, which will be used in combination with the relationships within the ontological structure of the domain.

It is important for the semantic analysis to retrieve the root of the word-form. We consider that stemming gives a result (that we call a ***Stem***) which is similar to the *root* or the basic semantic form of the word. We expect to use the stems to differentiate between different ontological categories. Because of this our further statistical analysis is based on the stems as discussed in the next section.

The more detailed global scheme of the system under development is given in Figure 4. The system is built of several modules (Python is mainly used) and a database (MySQL) for storing the word forms and the structure of the articles. As it is given on the scheme, Internet is parsed for news following a list of URL addresses. The retrieved articles are first kept in a text format for a short time for further classification and longtime storage in the database. As it is shown, the text itself is not stored, but the word forms are identified and stored in table word-forms, which is in fact the ED and contains for the moment 226658 word forms. The structures of the sentences, paragraphs and entire articles are stored in the corresponding tables. Additional information is extracted by the developed modules as follows:

1. The Part of Speech (POS) of each word form (one or more type of POS for the word-form);
2. The subject of the sentence is identified and stored in the table Sentence;
3. The word-form is stemmed and the normal form (stem) is stored in the table Roots which contains also the frequencies of the roots. The procedure of determining these frequencies will be explained in the next section;
4. The article is classified following the ontological categories of the domain under consideration.

One problem on which we concentrate in this paper is related to the classification of the articles to one of the 20 categories that are on the third level of the ontology we have developed (fig. 5).

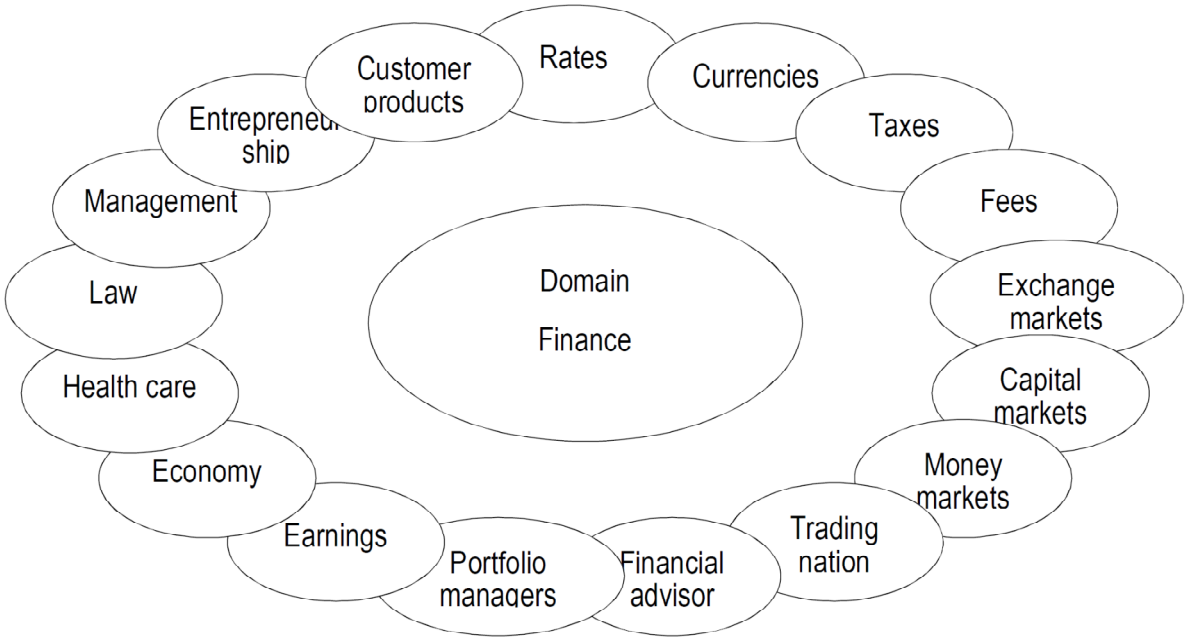


Figure 5. The third level of the Domain ontology developed for the purposes of the system

This problem of classification is related to the lack of explicit information on the Internet about the domain and the subdomain of a given article and, following from this, to the risk to store in the database articles that do not belong to the financial domain. We use this ontology information not only to identify the articles which are in the financial domain, but also because the further development of the system includes a knowledge representation module where the relationships between the events and the actors in the different branches of the ontology will be taken into account.

## Using heuristics for classification of articles to the ontological categories

The aim is to build a filter-classifier discarding the articles that do not correspond to the ontology of financial news developed for the system. Each article retrieved from the Internet has to be rejected or classified to one of the 20 third-level classes of the ontology (figure 5). For this purpose, we use statistical methods and heuristics. We assume that the developed tool has to work correctly without being time and memory consuming.

The main strategy is to develop a model of the most frequent word roots in each of the 20 categories and to check each newly retrieved article for belonging to these categories [Andonov & Slavova, 2014]. As a first step we have stored collections of 50 articles per category (see fig. 5), selected manually by human experts. Each category collection $C_i$ is further used as a global text model of the specifics of the language content for the category.

In one category $C_i$, each word form is stemmed and for each obtained stem $S_j$ the *category representation coefficient* $k_{ij}$ is calculated as follows:

$$k_{ij} = \frac{n_j}{n_{max}} . \log_2 \frac{1}{\frac{n_{jcor}}{n_{maxcor}} 2}$$

Where :

$n_j$ is the frequency of the stem j within the category text $C_i$

$n_{max}$ is the frequency of the stem with higher frequency in the category text $C_i$

$n_{jcor}$ is the frequency of stem j in the corpus Subtlex UK of English texts

$n_{maxcor}$ the frequency of the stem with higher frequency in the corpus Subtlex UK

We calculate the frequency $n_{jcor}$ of the stems by summation of the frequencies of all stem derivative word forms (given in Subtlex).

**The task of stemming**. Some of the extracted stems are for words which do not exist in the used corpus Subtlex UK, so we have enlarged the stems- dictionary (root) taken initially from the corpus. A separate module has been developed for this purpose (Fig. 4), which gets the words for which we cannot find the roots in the database. This is done by retrieving the word from Wikipedia and finding there its root which after that is also saved in the database.

The list of calculated $k_{ij}$ shows which stems are typical for the category $C_i$ and discards the stems which are frequently used in th**e** language in general. The list of first 50 most frequently met stems within a given ontological category $C_i$ represents its model $M_i$.

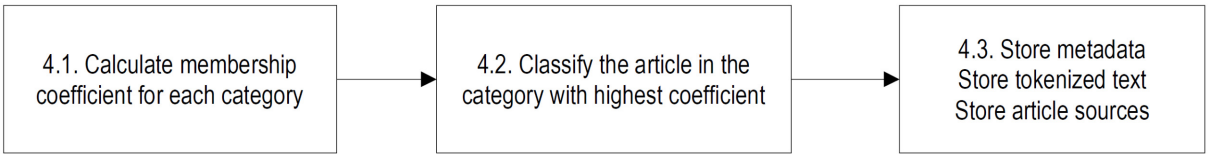


Figure 6. Heuristics classifier

Further the models $M_i$ (i=1 to 20) are used for classifying the new articles to the existing models $M_i$. Each novel article retrieved on the Internet undergoes a filtering-classification procedure based on the following formula which calculates its coefficient of membership $b_i$ to the category $C_i$.

$$b_i = \sum_{j=1}^{m} \frac{n_{stem\,j} \cdot k_{i\,stem\,j}}{n_{article}}$$

Where :

$n_{stem\,j}$ is the frequency of stem j (j = 1 to m) existing in the novel article

$k_{i\,stem\,j}$ is the category representation coefficient for the stem j in the model i

$n_{article}$ the number of stems in the novel article

$b_i$ is the membership coefficient of the novel article to the category Ci

As it is seen from the formula, the categorization is based on summation of the “influences” of stems in the newly arriving article multiplied by their corresponding category representation coefficient $k_{ij}$ for each category.

## Development of the system for opinion mining in a social context

As mentioned, the system is conceived to take into account the social opinion expressed about a news article. For that we retrieve posts in the social networks (SN) containing an URL to the article in question, assuming that the post expresses the attitude of its author toward the news. People with different social roles have different understandings and subjectivity level and for that we need to learn more about the person expressing the opinion. Our preliminary choice for the opinion holder’s characteristics are “gender“ and “level of education” as they are of importance in general. This initial set of characteristics will assist us in the methodology and technical approach for detecting more specific categories. In this paper we concentrate on the detection of the gender of the opinion holder.

The final goal is to obtain from the text and metadata of the posts a composite feature vector for classifying the gender categories and to use it in machine learning algorithm.

In order to analyze public reactions to financial (or any other kind of) news we concluded that the best way to harvest that data is to mine twitter. The reasons for that are that the twitter limit of 140 characters is actually beneficent for our task, that twitter is very popular and it has a great API.

Our first task is to identify the **gender of the person** who expresses the opinion. Unfortunately this attribute is completely lacking in the twitter metadata. There are several approaches of deducing the gender from the available information [Burger et al., 2011] – using the user name and the account name, using links in the user profile to social networks having gender attribute in their profiles, analyzing the text itself and using other cues.

**Using names.** Twitter API gives two attributes: name and screen_name. Screen name, being the account username, has a requirement to not include spaces, but apart from that in both fields the user can write basically anything she desires. Problematic values of those two fields include: symbols that are not characters; nicknames instead of real names; organizational accounts instead of a personal ones; names in any language and ethnicity; names that are not delimited with spaces or any other delimiting symbol.

We developed an algorithm for assigning a probability score of membership to male and female names by comparing substrings of the nickname with lists of known gender-specific personal names. The approach gave satisfactory results.

Another approach to twitter account gender identification is to link the user to one or more accounts in other social network services where gender information is available. Some twitter users use the URL and/or the description fields in their profile to post a link to their profile page in other social networks such as Google plus, blogger, LinkedIn, Facebook, etc. Initially we harvested twitter data for any web URL posted in the above mentioned fields. The observation of 98 974 tweets showed that, as a percentage, only about 3% of the users post any kind of URL in their profiles and those who do are usually merchant or organization accounts. Nevertheless, using users own gender expression is probably the most reliable method. As everywhere in the Internet the truthfulness of the information that users post online about themselves is not easily verifiable. It turned out however that some social networks are used for work and people there tend to use their real names, gender, etc. From all the major SN services, the most suitable for our needs turned out to be Google plus. It allows up to 20 million requests per day, which is more than enough for our needs. The drawback is that twitter users

who post a link to Google plus are very rare – out of 100 000 twitter posts less than 100 have Google plus URL. From those, not all have posted their gender, because the field is not required. As others did before us, we decided that this information can be used to create a training set for machine learning algorithm. The automatic classification should be based on commonly available features. in order to verify the gender identification made by other means. Using Google plus has other advantage though – it contains a lot of structured data about the user which can be helpful on a later stage.

In the application we developed we save all the extracted from different sources information about the user in a database. One record corresponds to one tweet and contains the meta-information about the user. From twitter we get the username, description, date, device, language, location and country and from Google plus - nickname, name, tagline, aboutMe, relationship status, occupation, organization, and places lived.

**Using tweet content.** The gender differences in the manner of writing were investigated in some recent studies and it has been shown that there is indeed a gender gap that cannot be accounted by external factors such as education, age, etc [Slik et al, 2015]. Others [Wassenburg et al, 2015] have found that girls construct more coherent and vivid mental simulations than boys and rely more heavily on these representations. From this we conclude that it should be possible to get information about the gender of the writer of some text by analysing the text itself. One way to do this is to use the “bag of words” approach. Unfortunately English language contains fewer gender cues than other languages. It still can be done by finding statistical differences between male and female word frequency usage. More complicated approach will be to analyze other text characteristics of the writing. Of course the two can be combined to create a more complex feature vector.

## Results

The initial tasks of downloading, classifying, storing and analyzing financial news texts are implemented and functional. The hypothesis that the stems can be used to perform the classification task is confirmed by experiments conducted with the system. The filtering procedure works correctly. The work done up to now encompass all the preliminary steps needed to be performed before the sentiment analysis should begin.

Concerning the social parameters of the opinion, there is no single approach for identifying author’s characteristics by written text that can give satisfactory results. Based on our work on the retrieved from

the SN data, we conclude that in order to identify the gender (and probably other characteristics) of a tweet's author, a complex feature vector is required that takes into account ***both*** the text itself and the available meta-information.

## Acknowledgments

This paper is published with partial support by the ITHEA ISS (www.ithea.org) and the Central Fund for Strategic development, New Bulgarian University.

This system has been developed at the New Bulgarian University and made part of a students' project for Internet databases development. We are glad to present our acknowledgments to Hani Akra, a student in computer science, from Syria, who accomplished the task related to the retrieval of new roots by consulting Wiktionary.

## References

[Andonov & Slavova, 2014] Andonov F & Slavova V. (2014) Some improvements of the OpenText Summarizer algorithm using heuristics, in proc. of the 10th Annual International Conference on Computer Science and Education in Computer Science.

[Burger et al., 2011] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (). Discriminating gender on Twitter. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, July. (pp. 1301-1309).

[Liu, 2010] Liu, Bing, (2010). "Sentiment Analysis and Subjectivity." Handbook of natural language processing 2 (2010): 568.

[Liu, 2012] Liu, Bing (2012). "Sentiment Analysis and Opinion Mining." Synthesis Lectures on Human Language Technologies 5.1 (2012): 1-167.

[Nicholls & Song, 2009] Nicholls, C. H. R. I. S., and Fei Song. "Improving sentiment analysis with part-of-speech weighting." Machine Learning and Cybernetics, 2009 International Conference on. Vol. 3. IEEE, 2009.

[Slavova and Hinkov, 2014] Slavova V, and B. Hinkov. Multimodal Sentiments Analyses of Financial News – a project outline, in proc. of the 10th Annual International Conference on Computer Science and Education in Computer Science, 2014.

[Slik et al, 2015] Van der Slik, F. W., Van Hout, R. W., Schepens, J. J. (). The Gender Gap in Second Language Acquisition: Gender Differences in the Acquisition of Dutch among Immigrants from 88 Countries with 49 Mother Tongues. PloS one, 10(11), e0142056, 2015. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0142056

[Wassenburg et al, 2015] Wassenburg, S. I., Koning, B. B., Vries, M. H., Boonstra, A. M., & Schoot, M. Gender differences in mental simulation during sentence and word processing. Journal of Research in Reading, 2015.

## Authors' Information

***Filip Andonov*** *- New Bulgarian University, department of Computer Science, fandonov@nbu.bg.*

***Major Fields of Scientific*** *Research: multi-criteria optimization, data mining, text processing, Python language*

***Velina Slavova*** *- New Bulgarian University, department of Computer Science, vslavova@nbu.bg*

***Major Fields of Scientific Research****: AI, Cognitive Science*

***Marouane Soula****, Ph. D Student at the Department of Computer Science, New Bulgarian University*