ARCHITECTURE OF A SENTIMENT ANALYSIS PLATFORM

CRISTIAN BUCUR
Lect. Dr.
University of Economic Studies, Bucharest, Romania
Petroleum And Gas University of Ploiesti, Romania
Email: cr.bucur@gmail.com

Abstract

A new domain of research evolved in the last decade, called sentiment analysis that tries to extract knowledge from opinionated text documents. The article presents an overview of the domain and present an architecture of a system that could perform sentiment analysis processes. Based on previous researches are presented two methods for performing classification and the results obtained.

Keywords: *sentiment analysis, opinion mining, supervised machine learning, unsupervised algorithms*

JEL Classification: *A12, M15, L21*

1. Overview

A sentiment analysis system could help process efficiently large volumes of information by identifying the opinion oriented phrases, and by eliminating subjective information. Also would be effective for summarization of opinions on large volumes of text.

Sentiment analysis can help determine if a product review is positive or negative, or if a customer is satisfied based on his posts online. Marketers can study how people respond to an ads campaign, or new products release. It could also help them determine how a product is seen by potential customers and identify the aspects that makes it good or not good in terms of sales.

Extracting the exact specifications of products people express opinions help obtain concrete data about how are considered the price, usability and viability, and where it is situated compared to the products of competition. Instead of using data from surveys made on customers, it helps collect valuable knowledge from consumers that are not company customers and help determine what might determine them to by buy a certain product.

Application using sentiment analysis could also be useful in areas like: politics, policy making, sociology and psychology. It could help analyze trends, determine how politician's messages are received by voters, and identify ideological bias. The process could help determine the attitude of people about their politicians, monitoring blog articles, or other platforms and the evolution of their attitude in time during a period of time (near or after elections). Sociologists could analyze the mass reaction on specific factors on a bigger scale relying also on a faster and cheaper method of collecting data than surveys.

2. Methodology

Sentiment analysis is defined as computational analysis of a text for classification and extraction of sentiments, opinion and feelings expressed by a subject [2]. An opinion or sentiment is defined like the following quintuple:

$$Opinion = < e_i, a_{ij}, oo_{ijkl}, h_k, t_l >$$

where h_k represents the emitent of opinion (the online user that express opinions regarding an entity), e_i represents the target entity of opinion (the entity on which the opinion is expressed by emitter, a_{ij} stands for an aspect of the entity,

oo_{ijkl} is the orientation of sentiment expressed by emitter h_k over an particular aspect a_{ij} of an entity e_i, and t_l represents the time when the opinion was expressed.

Having the following review:

Today ***I*** *got* ***Galaxy S5*** *phone, bought from online, yesterday and i noticed the quality of* ***materials*** *is built with. The* ***screen*** *quality is amazing, good* ***sound quality*** *and* ***battery life*** *seems pretty good.*

the emitter h_k represented by ***I,*** comments regarding the entity e_i, ***Galaxy S5***, over several aspects a_{ij}, ***materials***, ***screen***, ***sound, batter***. The expressed opinion orientation oo_{ijkl} at the moment t_l is mainly positive.

A sentiment analysis process could be made on several levels:

- Document level - the entire document is analyzed and are identified existing opinions and their polarity; Example of documents are posts from blogs or reviews for products.
- Sentences level - identify if there is an opinion expressed in a sentence and the polarity of that opinion
- Feature level - at this level are identified the attributes of the entity that are subject to opinion and the polarity of that opinion

In a sentiment analysis process, the main approaches are: machine learning approach and the approaches using a lexicon. Classification is made using supervised machine learning or semantic non supervised methods.

The machine learning approach treats the process as a text classification and uses several algorithms as support vector machine SVM, maximum entropy, k-nearest neighbor or Naive Bayes.

In ML (machine learning) methods, classifiers are built from annotated documents. In this case each document is considered a collection of words (bag-of-words). This is the method used by Pang, Lee and Vaithyanathan in a 2002 research. Other researches [1][7] include pre-processing of text utilizing natural language processing techniques for identifying features (stemming, lemmatization, stop word removal). This type of classifiers are domain dependent so their accuracy is lower in other domains than the ones they were trained in [5].

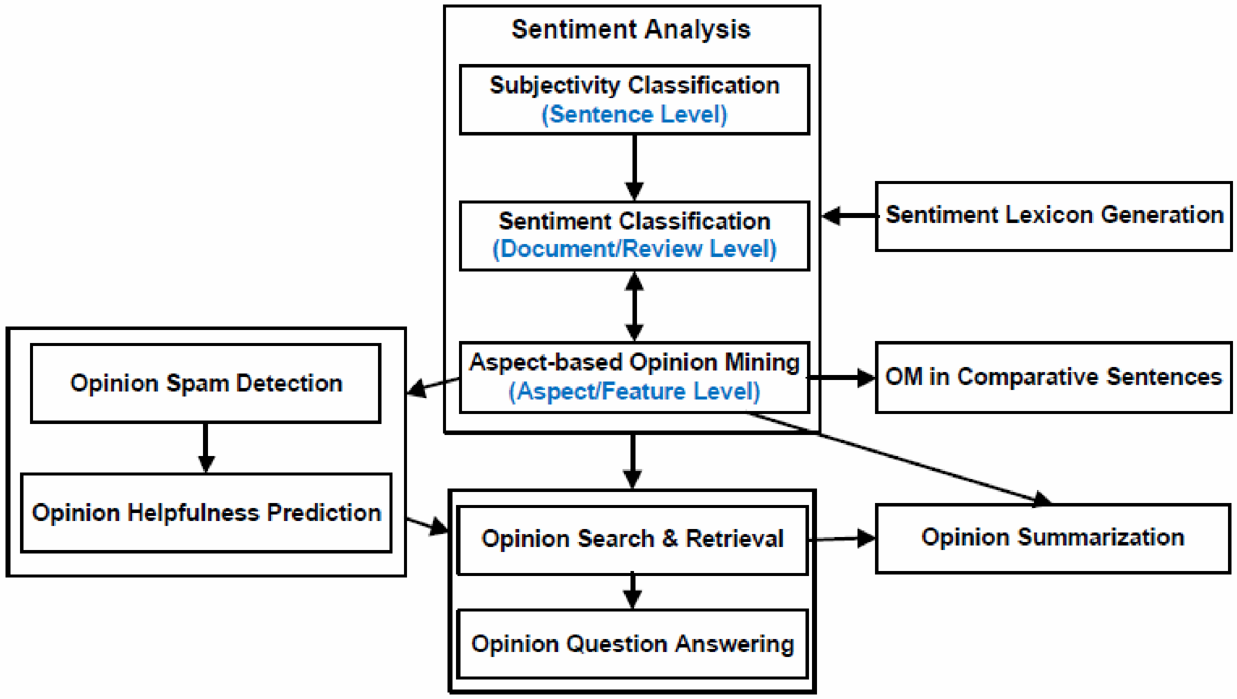
Sentiment Analysis
Subjectivity Classification
(Sentence Level)
Sentiment Classification
(Document/Review Level)
Aspect-based Opinion Mining
(Aspect/Feature Level)
Sentiment Lexicon Generation
OM in Comparative Sentences
Opinion Spam Detection
Opinion Helpfulness Prediction
Opinion Search & Retrieval
Opinion Question Answering
Opinion Summarization

Figure 1 Sentiment analysis tasks (source: http://iccke2014.um.ac.ir/uploading/iccke2014.um.ac.ir/images/ICCKE2014-eWorkshops-OpinionMining.pdf)

The unsupervised methods classifies documents using collections of sentiment words contained in evaluated text. One approach is to find patterns of words to determine the sentiment [6]. Turney in 2002 used a dictionary with words tagged with semantic orientation. The words sentiment orientation are summarized from their presents in documents using specific approaches [13].

Another approach is a dictionary based approach. Words sentiment is determined using lexicons with sentiment information.

There are several dictionaries used in sentiment analysis. WordNet is a lexicon that uses synonyms, antonyms and hierarchies to establish relations between words. Harvard General Inquirer (http://www.wjh.harvard.edu/~inquirer) attaches syntactic, semantic and pragmatic information to word tagged as part of speech. Linguistic Inquiry and Word Counts (LIWC) contains multiple categories in which are catalogued words that express emotions. Another lexicon used in recent researches in SentiWordNet. The lexicon is built automatically upon WordNet. Each sysnet from WordNet is annotated with three scores representing the degrees of positivity, negativity and objectivity.

Researchers also used hybrid approaches combining natural language processing and supervised machine learning. Nakagawa, Inui and Kurohashi [8], Zhang et al. [15]), Wu et al. [14] used approximated semantic properties with NLP as features in machine learning systems [5].

There are many challenges in conducting a sentiment analysis process.

- People have complex ways of expressing opinions
- Presence of irony and sarcasm
- The existence of negation
- Topic change in phrases
- There are situations in which a word could be considered positive and other situations in which the same word is considered negative
- Document expressing opinion contains mixed feelings; often people post reviews containing positive feeling regarding certain aspects and negative feelings for others
- The language used online is informal
- Documents could contain text in more than one language.

3. Sentiment analysis platform

A platform build for sentiment analysis should manage all the stages of the process. The platform should handle the acquisition of data, the store of information, the classification and the centralization of result. The best implementation solution is to use a modular platform.

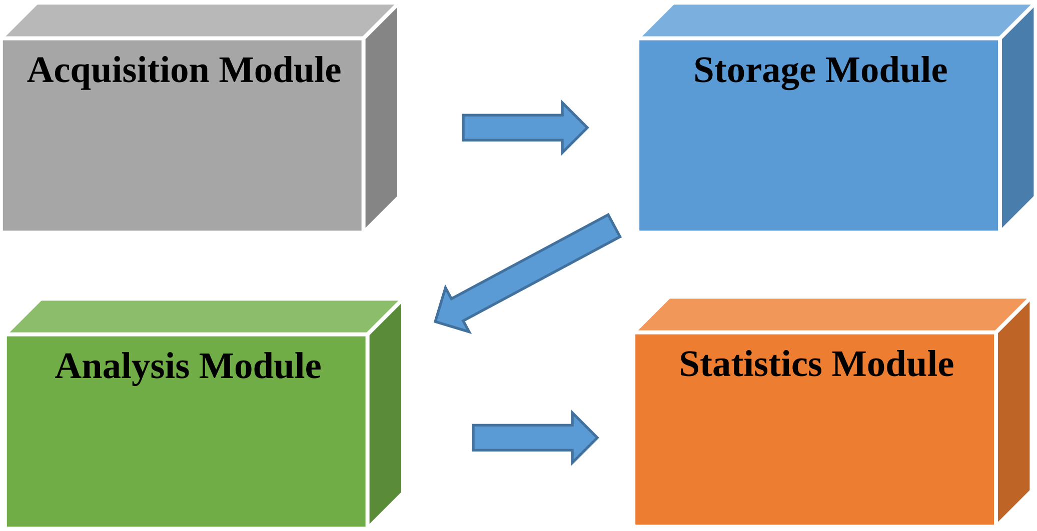

Figure 2 Sentiment analysis platform modules (source: author)

Platform has an Acquisition module that handles the collection of data from different sources. The module, extract data and keep track of the sources already processed. The Storage module handles the collected data used by analysis module. This usually consists from a database capable to store large volume of unstructured data. The platform database should be able to support multiple high speed read queries. This kind of specifications are specific to non-relational (no sql) databases.

The sentiment analysis process is handled by Analysis module. This is where the stored data is preprocessed and classified. As we described above the process could be done in multiple approaches.

In [3] we described a supervised algorithm made with a Naive Bayes classifier. Based on a pre-classified training set, the algorithm calculates the probability that a word has negative or positive meaning by the frequency of occurrence in each class.

The algorithm does not take into account the correlation between significance of context and word appearance but obtain good results with easy and quick implementation. For better accuracy the classification process necessitate pre-processing of data. The document is split into sentences and words. Each sentence is individually evaluated. It was determined that eliminating insignificant words (stop words) and calculating the probability of occurrence of groups of two words (2-grams) provide best results.

Another approach is using a lexical resource. As described in [4] we used an unsupervised process using SentiWordNet lexical resource. The data from storage module is also pre-processed. The evaluated document are split into sentences. Using a special algorithm we determined the part of speech from sentences (POS tagging). Then for each tagged word we determine the score according to the sysnet polarity from SentiWordNet. The sentence classification is established by calculating the scores for component words. According to SentiWordNet lexicon we obtained three scores, for objectivity, positivity and negativity. For a more accurate classification we used a threshold interval for determining the neutral sentences. Value lower than threshold were classified as negative and the higher ones as positive. The classification at document level was done by summarizing the sentences scores and using the threshold rule described.

The Statistics module of the platform is used for presenting the results obtained by the sentiment analysis process. This module could have an implementation for a user interface in which the result are graphically presented using chats. This module also implements and evaluation procedure for the algorithm precision. Usually the evaluation is done using natural language specific measures like: precision, recall, accuracy and f-measure like we described in [12].

For the two implementation described above we used the same collection of manually pre-classified reviews "sentence polarity dataset" (http://www.cs.cornell.edu/people/pabo/movie-review-data/) used by Pang and Lee in their research. This dataset contains movie reviews split into sentences collected from sites like imdb.com and rottentomatoes.com. There are over 5300 sentences classified as positive and negative.

For the supervised process in [3] we used more than 5000 sentences to train the algorithm. After data processing and optimizations for a set of 300 sentences we obtain the results presented below in table 1.

Table2. 1 Efficiency of supervised algorithm for groups of n=2 words

Precision	Recall	Accuracy	F Measure
0.82084690553746	0.76595744680851	**0.79939209726444**	0.79245283018868

(source: author)

In case of the lexicon based approach we used the same 300 set of sentences and obtained the following results (table 2):

Table. 2 Efficiency of unsupervised lexicon based algorithm

Precision	Recall	Accuracy	F Measure
0.57640750670241	0.65151515151515	**0.58636363636364**	0,61166429587482

(source: author)

We also compared the two algorithms implementation solutions in terms of volume of the data processed. In the chart below is presented the precision obtained on variable collection

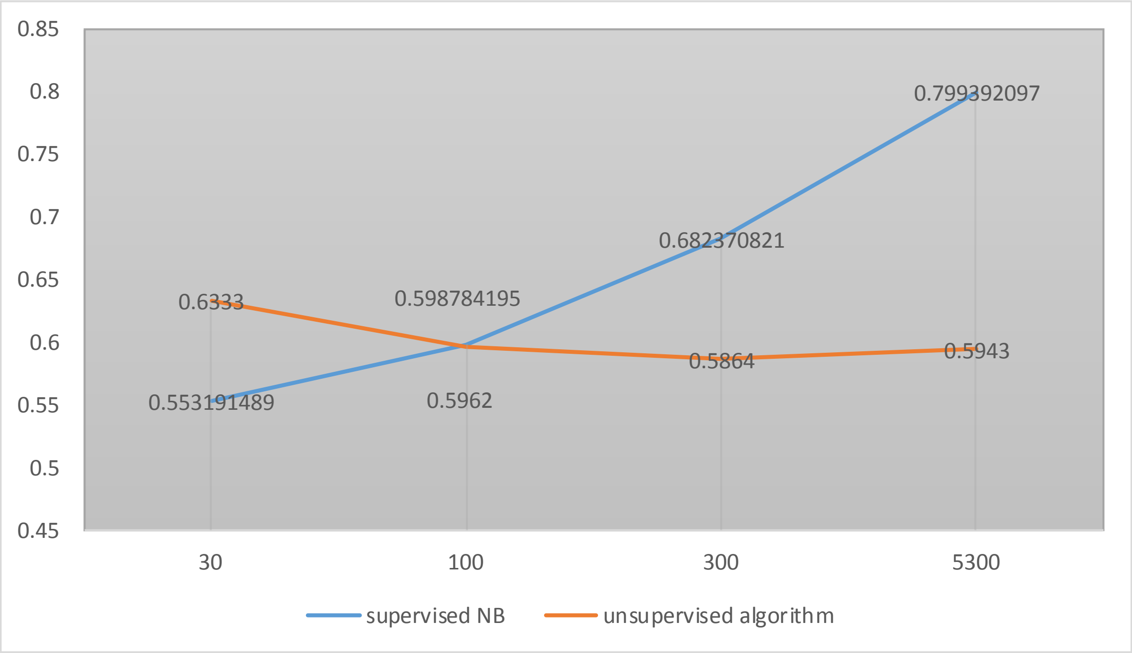

Figure 3 Comparison of precision obtained by presented methods

4. Conclusion

The article presents a review of sentiment analysis process and proposes an architecture of a platform capable of performing sentiment analysis. This platform has modular design and each of the module handle operation executed for acquire data preprocessing and classification. There are presented two methods for handling the classification process that could be implemented in acquisition module.

The two presented algorithms obtain different results in terms of precision. We could see that a supervised method is more precise in classification but necessitates a pre collected data manually classified for training the algorithm. Also the training dataset makes this process domain dependent.

The unsupervised lexicon based method is less accurate but require less effort not needing pre-classified data. This proposed method could be used in multiple domains with minor modifications.

From the results obtained we could observe the precision is influenced by training data in case of a supervised method while in case of lexicon based method the precision does not change.

5. Acknowledgements

This work was cofinanced from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/134197 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”.

6. References:

[1] Bakliwal, A., Arora, P., Madhappan, S., Kapre, N., Singh, M., and Varma, V. 2012. Mining sentiments from Tweets. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. WASSA ’12, Stroudsburg, PA, USA: Association for Computational Linguistics (pp. 11–18).

[2] Bing Liu (2011), Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data, Second Edition.: Springer, 2011.

[3] Bucur Cristian (2014a), "Aspects regarding detection of sentiment in web content", International Journal of Sustainable Economies Management (IJSEM), Volume 3: 4 Issues (2014), p.24-32, ISSN: 2160-9659

[4] Bucur Cristian (2014b), „Opinion mining platform for intelligence in business", Economic Insights – Trends and Challenges, Vol. III LXVI, No. 3/2014, ISSN 2284-8576 (http://www.upg-bulletin-se.ro/index.html#)

[5] David Vilares, Miguel A. Alonso And Carlos Gómez-Rodríguez (2015). A syntactic approach for opinion mining on Spanish reviews. Natural Language Engineering, 21, pp 139-163

[6] Mahmoud Othman, Hesham Hassan, Ramadan Moawad and Abeer El-Korany (2014). Opinion Mining and Sentimental Analysis Approaches: A Survey. Life Sci J 2014;11(4):321-326]. (ISSN:1097-8135)

[7] Montejo-Raez, A.,Martınez-Camara, E.,Martın-Valdivia,M. T., and Urena Lopez, L. A. 2012.Random walk weighting over sentiwordnet for sentiment polarity detection on Twitter. In Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. WASSA '12, Stroudsburg, PA, USA: Association for Computational Linguistics (pp. 3–10).

[8] Nakagawa, T., Inui, K., and Kurohashi, S. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In NAACL HLT'10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Proceedings of the Main Conference. HLT '10, Stroudsburg, PA, USA: Association for Computational Linguistics (pp. 786–94).

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

[10] Rahul Tejwani (2014), Sentiment Analysis: A Survey, http://arxiv.org/pdf/1405.2584.pdf

[11] Rodney Heisterberg,Alakh Verma, Creating Business Agility: How Convergence of Cloud, Social, Mobile, Video, and Big Data Enables Competitive Advantage, John Wiley & Sons, 2014

[12] Smeureanu, I., Bucur, C. (2012). Applying Supervised Opinion Mining Techniques on Online User Reviews. *Revista Informatică Economică, Vol. 16 No. 2/2012,*, 81-91.

[13] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. Lexicon-based methods for sentiment analysis. Computational Linguistics 37(2): 267–307.

[14] Wu, Y., Zhang, Q., Huang, X., and Wu, L. 2009. Phrase dependency parsing for opinion mining. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. EMNLP '09, vol. 3, Stroudsburg, PA, USA: Association for Computational Linguistics (pp. 1533–41).

[15] Zhang, C., Zeng, D., Li, J., Wang, F., and Zuo, W. 2009. Sentiment analysis of Chinese documents: from sentence to document level. Journal of the American Society for Information Science and Technology, 60(12): 2474–87.

[16] *** Introduction to Sentiment Analysis, Online, Retreived feb 2015,http://www.lct-master.org/files/MullenSentimentCourseSlides.pdf