Journal of Communication and Computer 14 (2017) 84-90
doi:10.17265/1548-7709/2017.02.005

Opinion Analysis on Web-based Reviews Using Support Vector Machine

Renato S. C. da Rocha[1], Marco Aurelio Pacheco[2] and Leonardo A. Forero Mendoza[3]

1. Department of Informatics, Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro 22451-900, Brazil

2. Electrical Engineering Department,Pontifical Catholic University of Rio deJaneiro, Rio de Janeiro 22451-900, Brazil

3. Electrical Engineering Department, State University of Rio de Janeiro, Rio de Janeiro 20550-900, Brazil

Abstract: This work aims to use sentiment analysis techniques, data mining, text mining and natural language processing to indicate the polarity of texts using SVM (support vector machine). Weka software and a movie review database from IMDb (internet movie database) were used. This work uses preprocessing filters and WRAPPER techniques and SVM for classification. It presents better results when compared to other preprocessing techniques used in sentiment analysis.

Key words: Text mining, sentiment analysis, machine learning, support vector machine, preprocessing techniques, filter, wrapper.

1. Introduction

Currently, the web user’s behavior is changing. As well as consuming content available on the web, people are also exposing their opinions and experiences about products acquired, places they visited and services they used. These reports may influence the decision of other users, serving as additional information that is not often available in the description of the product. This feedback process can be very useful both for companies, which may use this information to improve their products, and for consumers, who can also take advantage of the experience of other users.

Sentiment analysis area is dedicated to the analysis of opinions, feelings, emotions and attitudes of people and has become an object of research in various scenarios from social media such as blogs, e-mail, discussion forums, products review, etc. Opinion mining is a part of text mining with focus on processing user-generated content. This feature adds several research challenges such as identifying topics and opinions. User-generated content is considered unstructured data, as they may deal with various subjects in the form of free text. In order to make the data more easily understandable, developed methods seek to process the opinions so that they can be represented in a structured way. Structured data allow a straightforward analysis of the main topics along with the given average opinion. With the growing popularity of social media, opinions processing tools have to deal with large amounts of data. Thus, it becomes imperative to represent data in a summarized form. The task of sentiment analysis is quite complex, firstly by language problems inherent in the problem, such as sentences formation and spelling rules. Secondly, depending on the application, for example, how to classify movie reviews, there are people who use irony and sarcasm, which should be identified by the algorithms. Thirdly is the data preprocessing and structuring. In this regard, classical techniques are applied, but rely heavily on the specialist’s knowledge.

This paper presented a classification of a movie review database, using SVMs (support vector machines) as the classifier algorithm. For data preprocessing, classical techniques presented in various works were compared with a mix of classical

Corresponding authors: Renato S. C. da Rocha, M.Sc. student, research field: machine learning.

filters and the wrapper preprocessing technique, which performed better in addition to avoiding specialist interference.

2. Fundamentals of Text Mining

Text mining is an important step of KDT (knowledge discovery from text) [1]. According to Ref. [2], text mining is a variation on data mining: while regular data mining extracts patterns, regularities or other trends from structured databases of facts, text mining leads with problem of natural language processing.

Gleaning useful information from natural language text, however, has been a daunting task because text is amorphous and unstructured. Unstructured information, which mostly originates from social media, constitutes 80% of the data worldwide and accounts for 90% of Big Data [3].

Most unstructured data are not modeled, are random, and are difficult to analyze. In particular, text is far more complex, involving cultural nuances in the communication of information, opinions, or dramatic narrative. Nonetheless, with the advancements in data storage and the ready availability of digitized texts, text mining can help in acquiring better competitive intelligence.

Text mining uses techniques from information retrieval, NLP (natural language processing), statistics, machine learning, and specially data mining. The choice of the techniques depends on the dataset and the nature of text mining task. Typical text mining tasks include text classification, text clustering, information extraction, and sentimental analysis [4].

This work has been financially supported by CNPq and FAPERJ.

Sentiment analysis (also called opinion mining, review mining, appraisal extraction or attitude analysis) is the task of detecting, extracting and classifying opinions, sentiments and attitudes concerning different topics, as expressed in textual input [5, 6].

The next subsections provide a brief introduction to opinion mining, as well as to the SVM classifier that constitutes the essence of this work.

2.1 Sentiment Analysis

The sentiment analysis is the computational study of expressed opinions, often subjective, in any snippet of text in natural language (document). The entire opinion is composed of at least two basic elements: a target and a sentiment about the target [7]. A target consists of an object, also called entity or topic, or characteristics of the object, also called aspects, which can be a product, person, organization, brand, event, among others. The terms sentiment and opinion are often used synonymously in this context. The polarity of a sentiment may be classified into discrete classes (positive, negative or neutral) [8], or as a range that represents the intensity of the sentiment, typically [-1, 1]. Besides entity, aspect and polarity, the other two characteristics that complete an opinion quintuple [7] are the opinion source and the time the opinion was expressed.

Thus, the main sentiment analysis task can be defined as follows: given a document D, identify the expressed opinions about an entity, its aspects and polarity. A document can be analyzed at different levels of granularity: i) the lower the granularity, the more specific the classification; ii) the decision level is subjected to the context and application. In this sense, an opinion can be classified as to its polarity in terms of: document, sentence, entities and aspects.

Opinions may be regular or comparative; direct or indirect, and implicit or explicit [7]. The way opinions are expressed directly influences the ability to properly process them. Because they are easier to be treated, most works focus on regular, direct and explicit opinions. The challenges in the opinions processing, inherent in natural language processing, are: words disambiguation; sarcasm and irony; semantics and syntax, among others.

Some steps are necessary to perform the sentiment analysis, since text mining originates from multiple sources in various formats. In general, a process of text mining occurs in five macro steps [9]: data collecting, preprocessing, indexing, mining and analysis.

The first step aims at gathering information to compose the textual database to work (corpus), i.e., it involves determining and selecting the universe to which the text mining techniques will be applied. Collecting social data is usually done with APIs (application programming interface).

After collecting documents, the preprocessing step structures them for the automatic knowledge extraction algorithms application. Primordial to the entire mining process performance, the preprocessing operations include: sectioning of a document in minimum units with the original text semantics (tokens) and removal of tokens without semantic and irrelevant value for mining (stop words) [10].

There is also the possibility of applying statistical information based filters that influence the classification quality: IDFTransform method (Inverse Document Frequency Transform) takes the premise that the attributes that rarely appear are valuable for classification, while the TFTransform method (Term Frequency Transform) admits that the most common terms are more important. Additionally, the following NLP techniques are often used, improving results: order and position of words identification, grammatical classes labeling, speech analysis, reduction of derived words to their root form (stemming), and the conversion/correction of informal writing, abbreviations and emphasis on words by repeating characters, quite common in social networks and that produce inaccurate evaluations by traditional mining techniques.

Thereafter starts the indexing phase. Indexing is the process responsible for creating auxiliary structures called indexes that guarantee speed and agility in the recovery of documents and its terms. Two more efficient distinct approaches are present in the text mining works: textual indexing and thematic indexing [11].

Once indexed, documents and terms are subjected to machine learning algorithms to perform knowledge extraction (mining step).

Finished the mining step, the sentiment analysis of extracted messages is carried out. The goal is the positive, negative or neutral polarity classification.

2.2 Polarity Classification

Literature divides different sentiment classification techniques on three approaches: lexicon based approach, machine learning approach and hybrid approach [12].

The lexicon approach is based on a collection of sentiment terms previously known and pre-compiled and can be of two types: dictionary-based [13] and corpus-based [7]. A dictionary consists of a database comprising words previously classified according to their polarities, and can be constructed either manually or from other words, called seed words. In the corpus-based approach, the semantics technique is very similar to the statistical, except the polarity is measured in terms of some measure of distance between terms, often PMI (pointwise mutual information) [14]. The techniques principle in this category is that semantically close words must have the same polarity.

In machine learning-based approach, supervised methods are employed, classification to be more specific. Basically, these methods consist of the execution of two processes: i) learn a classification model on a training corpus with previously labeled classes (positive and negative, for example); ii) use the model obtained in i) to classify documents that were not used in the construction of the classifier. SVM is among the most successful algorithms in classification tasks [15].

SVM algorithm represents documents as points in a vector space, which dimensions are selected features. Using the document training vectors, the basic idea of SVM is to find the optimal hyperplane that separates the previously classified data with the largest margin of separation between the two classes. The optimal hyperplane is then used to classify unlabeled data. The support vectors are those that define the optimal hyperplane separation location. SVMs deal, very effectively with non-linear problems, mapping the training set of its original space to a new larger space, outperforming other techniques such as artificial neural networks [16]. Literature describes a wide range of SVM application in text mining tasks [6].

However, in high-dimensional feature space, supervised methods, such as SVMs, suffer due to the curse of dimensionality [17]. A possible solution to avoid this issue is to use feature selection techniques. They are often used to reduce the dimensionality of the feature space and improve computational efficiency and accuracy of classifiers [18]. One successful approach for feature selection, which fits very well the Web data extraction problem [19], is based on wrappers [20]. Wrappers search for an optimal feature subset using the classification accuracy of some learning algorithm as their evaluation function. Thus, the best search-fit is an optimization problem and, therefore, several techniques can be used, including the evolutionary algorithms. Evolutionary algorithms, such as genetic algorithms [21], are population- based metaheuristics of great research interest because of their promising results in different application. These metaheuristics use principles of Darwin's theory of natural selection: at each generation or iteration of the algorithm, a competitive selection occurs to choose the best solutions; these are modified by crossover and mutation operators to generate new solutions, repeating this cycle until a given stop criterion, defined by the user, is reached.

The next section describes the proposed method for classifiers construction.

3. Methodology

The sentiment classification task will be divided into three steps: information collecting is the first step, database preprocessing, the second step, and the third and last step is the database classification to find polarity of each test observation. Weka software was used to develop this work.

3.1 Database

An existing database was used, and the information collecting has been previously made. The database named "newsgroup rec. arts. movies. Reviews" from IMDb Internet Movie Database [5] contains several movie reviews, which were collected, classified between positive and negative and made available to test sentiment analysis algorithms. This database is quite complex from the sense of SA (sentiment analysis) since movie reviews contain ironies and sarcasms which can affect the performance of classification algorithms. The database consists of two thousand files divided into two groups: 1,000 observations with positive polarity and 1,000 observations with negative polarity.

Being a known and widely used benchmark, this database facilitates results comparison with other algorithms.

3.2 Data Preprocessing

Firstly, the database was divided by tokens. Those tokens will be the text's words. Then, three different preprocessing configurations were used to compare the classification result and determine which technique is more efficient.

(a) The first preprocessing methodology used the following techniques:

- IDFTransform;
- LowerCaseTokens;
- MinTermFreq;
- StopWords;
- Stemmer.

Fig. 1 shows the steps to the first preprocessing methodology.

(b) A genetic algorithm wrapper was used in the second preprocessing methodology. The SVM classifier was the evaluation function and the classification accuracy was used to evaluate the generated solutions.

The GA (genetic algorithm) used the database set of words as the basic chromosome; each chromosome gene comprises a database word. Genetic mutation and crossover operators were used, with fixed rates of 0.3 and 0.6, respectively.

Fig. 2 shows the steps to the second preprocessing methodology.

(c) In the third preprocessing methodology, both filters presented on the first and wrapper presented on the second methodology were used.

Fig. 3 shows the steps to the third preprocessing methodology.

3.3 Database Classification

For each preprocessing methodology, a SVM classifier with two different kernels was used: polynomial and RBF (radial basis function). Experiments were done with different settings until the best configuration was reached. It used 80% of the database for training and 20% for testing.

4. Results

In Section 3.2.a preprocessing methodology was used in the first experiment. The trained SVM model was tested with the polynomial kernels and the radial basis function. The values of the exponent and the complex variable C are changed in the polynomial kernel. The values of o and the complex variable are changed in the RBF kernel. Results are presented below in Tables 1 and 2:

The best classification configuration with the first preprocessing configuration was obtained with exponent 0.1 and C = 2.0 with polynomial kernel, resulting in 89.3% accuracy.

The second experiment of Section 3.2.b, preprocessing methodology was used. The trained SVM model was tested with the polynomial kernels and the radial basis function. The values of the exponent and the complex variable C are changed in the polynomial kernel. The values of o and the complex variable are changed in the RBF kernel. Results are presented below in Tables 3 and 4.

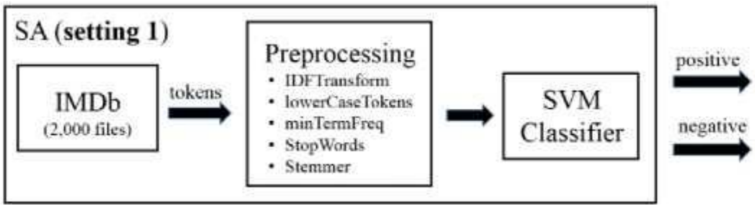

Fig. 1 First preprocessing methodology.

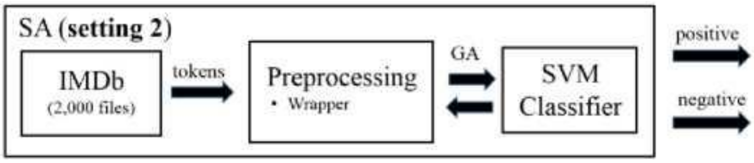

Fig. 2 Second preprocessing methodology.

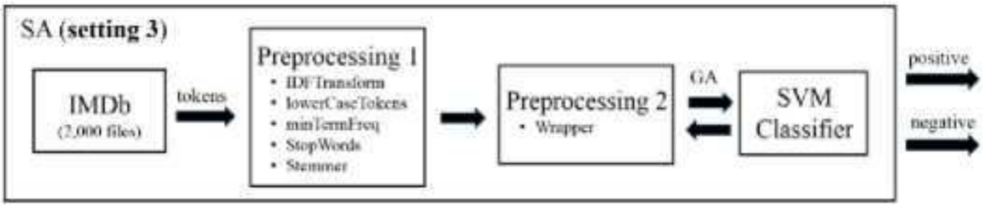

Fig. 3 Third preprocessing methodology.

Table 1 RBF Kernel models for first preprocessing methodology.

RBF Kernel				
	g = 0.1	g=0.05	g=0.01	g=0.005
c = 1.0	48.6%	71.3%	87.6%	87.3%
c = 1.5	50.6%	73.3%	86.6%	87.6%
c = 2.0	50.3%	73.3%	86.0%	86.0%

Table 2 Polynomial Kernel models for first preprocessing methodology.

Polynomial Kernel				
	e = 1.0	e = 0.5	e= 0.25	e= 0.1
c = 1.0	78.6%	82.0%	86.0%	88.6%
c = 1.5	77.6%	83.0%	85.0%	89.0%
c = 2.0	77.6%	81.0%	83.0%	89.3%

Table 3 RBF Kernel models for second preprocessing methodology.

RBF Kernel				
	g=0.1	g=0.05	g=0.01	g=0.005
c = 1.0	49.4%	74.3%	88.6%	89.3%
c = 1.5	57.1%	75.3%	89.6%	89.6%
c = 2.0	57.2%	75.3%	86.0%	90.2%

Table 4 Polynomial Kernel models for second preprocessing methodology.

Polynomial Kernel				
	e = 1.0	e = 0.5	e = 0.25	e= 0.1
c = 1.0	80.8%	84.0%	88.0%	88.9%
c = 1.5	77.6%	85.2%	91.1%	88.9%
c = 2.0	77.6%	85.2%	91.1%	88.9%

Table 5 RBF Kernel models for third preprocessing methodology.

RBF Kernel				
	g=0.1	g=0.05	g=0.01	g=0.005
c = 1.0	61.4%	77.3%	90.4%	90.4%
c = 1.5	62.8%	78.9%	90.4%	91.6%
c = 2.0	62.8%	78.9%	89.8%	91.6%

Table 6 Polynomial Kernel models for third preprocessing methodology.

Polynomial Kernel				
	e = 1.0	e = 0.5	e= 0.25	e= 0.1
c = 1.0	88.3%	88.2%	90.1%	93.7%
c = 1.5	87.9%	88.7%	92.6%	93.7%
c = 2.0	87.6%	88.9%	92.6%	92.6%

The best classification configuration with the second preprocessing configuration was obtained with exponent being 0.25 and C = 1.5 with polynomial kernel, resulting in 91.1% accuracy.

In the third experiment of Section 3.2.c preprocessing methodology was used. The trained SVM model was tested with the polynomial kernels and the radial basis function. The values of the exponent and the complex variable C are changed in the polynomial kernel. The values of o and the complex variable are changed in the RBF kernel. Results are presented below in Tables 5 and 6.

The best classification configuration with the second preprocessing configuration was obtained with exponent being 0.1 and C = 1.5 with polynomial kernel, resulting in 93.7% accuracy.

5. Results Analysis

The third preprocessing configuration, which blends classical techniques of text mining approach to the wrapper, outperformed the other two models. This hybrid data preprocessing methodology was very efficient, and had better accuracy in almost every SVM configuration when compared with the other preprocessing methodologies.

The method shown in the second part of the results section shows how wrapper technique alone achieves better accuracy than the classic filter methods of text mining shown in the first part of results.

These results were compared with two other sentiment analysis classification works: Ref. [22] achieved 90.3% and Ref. [23] achieved 81%, both working with deep learning techniques, using the same database. The best result achieved in this work 93.7% performed better than this other two works.

By using the genetic algorithm wrapper methodology, this model has a slightly higher computational cost when compared to deep learning techniques, but the results are better and consistent. As it is an optimized model, it can be more robust to changes in the database than traditional methods.

6. Conclusion

This paper has presented a sentimental analysis model, combining the wrapper method with the SVM classifier. This model has improved the text classification compared to other models using the same database as test.

The results show the wrapper-preprocessing filter can effectively clean the data. When it is used jointly with classical preprocessing filters, it provides superior results. This technique is not found in sentimental analysis tasks, but it is often used in data mining. The classification task heavily depends on data cleaning. As stop words and stemmer are very subjective, the wrapper suffers less influence from specialist, being more robust.

References

[1] Hotho, A., Numberger, A., and PaaB, G. 2005. “A Brief Survey of Text Mining.” *LDV Forum—Gld. J. Comput. Linguist. Lang. Technol.* 20: 1962.

[2] Hearst, M. A. 2003. “What Is Text Mining?” [Online]. Available:http://people.ischool.berkeley.edu/hearst/text-mining.html. [Accessed: 12-May-2016].

[3] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M., and Gani, A. 2014. “Big Data: Survey, Technologies, Opportunities, and Challenges.” *Sci. World J.* 2014: 1-18.

[4] Feldman, R., and Sanger, J. 2006. *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press.

[5] Montoyo, A., Martinez, B. P., and Balahur, A. 2012. “Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments.” *Decis. Support Syst.* 53 (Nov.): 675-9.

[6] Ravi, K., and Ravi, V. 2015. “A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications.” *Knowledge-Based Syst.* 89 (Nov.): 14-46.

[7] Liu, B. 2012. “Sentiment Analysis and Opinion Mining.” *Synth. Lect. Hum. Lang. Technol* .5 (May): 1-167.

[8] Tsytsarau, M., and Palpanas, T. 2012. “Survey on Mining Subjective Data on the Web.” *Data Min. Knowl. Discov.* 24 (May): 478-514.

[9] Aranha, C. N.2007. “An Automatic Preprocessing for Text Mining in Portuguese: A Computer-Aided Approach.” Pontifical University Catholic of Rio de Janeiro, Brazil.

[10] Konchady, M. 2006. *Text Mining Application Programming*, 1st ed. Rockland, MA, USA: Charles River Media, Inc.

[11] Yates, R. B., and Neto, B. R. 2011. *Modern/n/ormation Retrieval: The Concepts and Technology behind Search* (2nd Edition) (ACM Press Books). Addison-Wesley Professional.

[12] Medhat, W., Hassan, A., and Korashy, H. 2014. “Sentiment Analysis Algorithms and Applications: A Survey.” *Ain Shams Eng. J.* 5 (Dec.): 1093-113.

[13] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. “Lexicon-Based Methods for Sentiment Analysis.” *Comput. Linguist.* 37 (Jun.): 267-307.

[14] Church, K. W., and Hanks, P. 1989. “Word Association Norms, Mutual Information and Lexicography.” In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*.

[15] Vapnik, V. N. 2000. *The Nature of Statistical Learning Theory*. New York, NY: Springer.

[16] Mendoza, L. A. F. 2009. “Redes Neurais e Maquinas de Vetores de Suporte no Reconhecimento de Locutor usando Coeficientes MFC e Caracteristicas do Sinal Glotal.” Universidade Federal Fluminense, Rio de Janeiro, Brazil (in Portuguese).

[17] Bellman, R. 1957. *Dynamic Programming.* 1st ed. Princeton, NJ, USA: Princeton University Press.

[18] Chen, J., Huang, H., Tian, S., and Qu, Y. 2009. “Feature Selection for Text Classification with Naive Bayes.” *Expert Syst. Appl* .36 (Apr.): 5432-35.

[19] Ferrara, E., DeMeo, P., Fiumara, G., and Baumgartner, R. 2014. “Web Data Extraction, Applications and Techniques: A Survey.” *Knowledge-Based Syst.* 70 (Nov.): 301-23.

[20] John, G. H., Kohavi, R., and Pfleger, K. 1994. “Irrelevant Features and the Subset Selection Problem.” In *Machine Learning Proceedings 1994,* Elsevier.

[21] Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-We.

[22] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. 2011. “Learning Word Vectors for Sentiment Analysis.” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142-50.

[23] Martineau, J., and Finin, T. 2009. “Delta TFIDF: An Improved Feature Space for Sentiment Analysis.” In *Proceedings of the Third AAA/ /International Conference on Weblogs and Social Media*.