World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:3, 2014

Contextual Sentiment Analysis with Untrained Annotators

Lucas A. Silva, Carla R. Aguiar

Open Science Index, Computer and Information Engineering Vol:8, No:3, 2014 publications.waset.org/9997689.pdf

***Abstract*—**This work presents a proposal to perform contextual sentiment analysis using a supervised learning algorithm and disregarding the extensive training of annotators. To achieve this goal, a web platform was developed to perform the entire procedure outlined in this paper. The main contribution of the pipeline described in this article is to simplify and automate the annotation process through a system of analysis of congruence between the notes. This ensured satisfactory results even without using specialized annotators in the context of the research, avoiding the generation of biased training data for the classifiers. For this, a case study was conducted in a blog of entrepreneurship. The experimental results were consistent with the literature related annotation using formalized process with experts.

***Keywords*—**Contextualized classifier, naïve Bayes, sentiment analysis, untrained annotators.

I. Introduction

According to Internet World Stats [1], the internet usage has grown more than 500% between years 2000 and 2012. Along with that came web 2.0, which gave a power of unprecedented sharing of information and opinions to Internet users. These facts, today, can be represented by the large scale usage of social media and blogs, which, as shown by Nielsen Company [7], dominate about 10% of the time spent on the Internet. Such user-generated content has been analyzed from a social and content-oriented point of view [20]. The social network analysis techniques can work, for example, identifying the polarity of opinions in user comments for rating movies [19]. This technique is called sentiment analysis which is formally defined by Liu [13] as the evaluation of opinions, assessments, attitudes and emotions before entities, individuals, specific topics, events and their attributes which indicate positive or negative feelings. Sentiment analysis finds use in various sectors of the consumer market, such as product evaluation, discovery of their attitudes and consumer trends to strengthen marketing campaigns, find opinions about hot topics or review movies.

Within the context of use of the sentiment analysis, there are a large number of opinions available, and since these have to be labeled (such as negative, positive or neutral). Given this, is true that a solution to optimize sentiment analysis is to use machine learning techniques, more specifically, supervised learning algorithms. Remember that to use supervised learning techniques is necessary a corpus consisting of manually annotated sentences for a given context. This process is usually carried out by trained annotators in the context analyzed, as can be seen in Wilson [27]. However, one factor to be emphasized here is that the time to train annotators can be high (Wilson [27] used a period of four years).

This paper attempts to validate the hypothesis that it is possible to increase the efficiency of a classifier for sentiment analysis from data provided by untrained annotators that produce inputs contextualized. In order to achieve this goal a web platform, denominated "*Analisador de Sentimentos*" (Sentiment Analyzer), was developed to perform the entire case of study with entrepreneurship blog to assess the trend that blogs authors of article have to express sentiment in their posts. The procedure followed can be described with four main steps:

1) Collect data
2) Initial evaluation of blogs posts
3) Creation of a corpus using untrained annotators
4) Final evaluation of blogs posts

The remainder of the paper is organized as follows. Section II will introduce related works in sentiment analysis. Section III will describe the methodology used, covering the web platforms pipeline of execution and the automated process of annotation. Section IV will discuss the results obtained. Finally, Section V will conclude this paper.

II. Related Work

This section briefly surveys previous work on sentiment analysis.

One area of research is the evaluation of reviews of movies or products. In this context it is possible to find works such as Kim and Hovy [14], which used a maximum entropy model and features with lexical structures and position to extract pros and cons from online reviews. Liu et al. [12] present a system for automatic summarization of reviews. They capture the characteristics of the products and identify whether the sentences are positive or negative about them. A similar research could be found in Kushal [8].

Some techniques also take into account a set of words with the semantic weight and use a dictionary, like WordNet [5]. New words generated from the synonyms found in dictionary. Liu et al. [12] used this technique. This idea of using a small set of words and expand vocabulary through a dictionary was suggested by Turney [22]. A similar work can be found in Bergler [6].

In general, machine learning techniques are fairly widespread in sentiment analysis. Researchers Pang and Lee

L. A. Silva is with the Department of Computer Science, University of Brasilia, DF, Brazil, Campus Darcy Ribeiro (e-mail: loamilucas@gmail.com).

C. R. Aguiar is with the Department of Software Engineering, University of Brasilia, DF, Brazil, Campus Gama (e-mail: rocha.carla@gmail.com).

[19] make a comparison among three different algorithms: maximum entropy [16], support vector machine (SVM) [9] and Naive-Bayes, using a diverse set of features. Among the main contributions of this work are finding that feature presence is more important than feature frequency and improves accuracy if all the frequently occurring words from all parts of speech are taken.

There are also approaches for unsupervised algorithms. Turney [21] seeks to identify the semantic orientation of a phrase using PMI-IR. Guidance is given by comparing the words with a positive reference, the word “excellent”, and a negative reference, the word ”poor”. The mutual information is computed using statistics gathered by the search engine. Importantly, as the search engine used by the researcher is no longer available, it is difficult to replicate this technique.

It is important to mention researches developed using annotators to produce labeled information. Within this context, we find Wiebe [23], [26] that describes a process of manual annotation of opinions, emotions, sentiments, speculations, evaluations and other private states in language. Bermingham [17] built a corpus based on capturing information from web and removes the noise. He used 7 annotators to label the captured data. He also conducted a sentiment analysis binary (i.e., classifying sentences into positive or negative) and ternary (adding the neutral category). The researcher obtained better results with binary classification. The work that probably comes closest to the research presented in this paper is from Wilson [27]. The researcher used formal training method for annotators, where they passed two levels of training, one to learn how to manually annotate to annotate and the second using the tool GATE. Her research presents data as identification of anchors in text, differences between objective and subjective annotations of texts and identification of intensity of private states.

III. Methodology

As described in the introduction, the methodology applied in this paper presents four steps: data collection, the initial evaluation of the data, creating a new database and new classifiers and a final evaluation of the results presented by all classifiers generated. The first step consists in capturing the contents of the entrepreneurship blog to be studied and storing this in the database of the Sentiment Analyzer. The second stage receives the content captured in the previous and performs inferences about the polarity of all the sentences. This activity is done with a generic classifier. All assessment performed is stored in the database. These first two steps represent a macro process called initial data classification which will be described subsequently.

Later, we selected a fraction of the collected database to be analyzed by untrained annotators, which produce relevant inputs to the generation of files that will feed the new classifiers. After this step, all the new classifiers and the generic classifier will be used to analyze the training base so that their accuracies are compared and studied. These steps constitute the second macro process called annotation and final data classification.

A. Initial Data Classification

This process aims to get the blog's articles and provide classification for them using a default classifier, trained with a non-contextualized corpus (Fig. 1). The capture of blog data was performed using the web scraper module of the Sentiment Analyzer. The web crawler acts in a non-recursive way in blog pages. It takes as input the root URL and the number of pages that must be captured and produces as output a set of URLs of the blog articles section for Startups. The web scraper extracts the articles content that exists in the URLs gotten by crawler. Importantly, this component performs cleaning filters in text, which removes images, videos, javascript codes and advertising inserted within the text. The data captured in step described will be used throughout the study performed in this research.

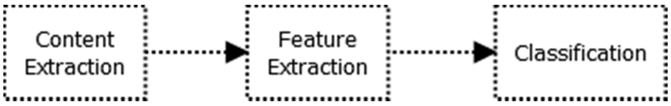

Fig. 1 Scheme for first process of the Sentiment Analyzer

To select the features that would be evaluated by the classifier, a study was done with a few basic phrases collected. This study aimed to identify the factors that impact on the polarity of the sentences. The six sentences below will be used as examples. They were extracted from the generated database of articles.

1) "Skit is an iOS app that allows you to import images [...] on the Internet and string them together into fun little animated cut-out movies [...]"
2) "We’re thrilled to have him on the team."
3) "But the data behind an experiment can be so messy"
4) "In a statement, Tom Impallomeni, CEO of Swapit, adds: 'With SuperAwesome we’ve created the biggest kids and teens discovery platform in the UK which is safe, compliant and effective' "
5) "Until we can zap Bitcoins to the cashier at Arby's, we're not really living in the future"
6) "And yes, the writing on English-language blogs can be pretty rough, too"

The division of sentences into elements of sentiment analysis [13] shows that all keywords are adverbs, adjectives and some prepositions [10], [11], [24], [25]. It is noteworthy that these words have a size greater than or equal to three letters. Entities are represented by a noun or a pronoun. It is also noticeable that elements such as articles or prepositions do not contribute to the classification of the polarity of the sentences, appearing only as connectors within the text.

Given this context, it is plausible to use unigrams [15], [18] that have size (number of letters) greater than or equal to three as features. Thus, the general structure features are:

For example, for the phrase “We’ve thrilled to have him on team.” the set F of features is:

This set of features has the limitation that doesn't enhance the elements that determine the polarity. It only removes those that certainly not contribute to polarity.

The last stage of the algorithm uses Naïve Bayes, to perform the classification of sentences. The Naïve Bayes classifier [3], [4], [13] previously trained to carry out the initial classification of the sentences of the blog. For this, we applied the Cornell University's dataset, called "sentence polarity dataset" [2], which contains sentences with positive and negative polarities and there is no neutral sentences. It is worth highlighting that this basis is not contextualized with the topic of startups, so it's a generic basis. The classifier generated in this process will be called generic classifier. It was generated so that, in the experiment to be described later, the results are compared with those of the new classifiers created.

B. Annotation and Final Data Classification

The second process aims to classify blogs articles using a contextualized classifier, trained with a contextualized corpus, created from the data collected (Fig. 2). It starts with a selection of non-classified articles from database and then these articles are applied to annotators, which can give to each articles phrases one of these three sentiment category: positive, negative and neutral.

The annotators are a key part within the second macro process. Their importance is given to the extent that is necessary to label the sentences collected to develop the learning algorithm. The better the annotation process, i.e. the more statements are correctly classified, the classifier will be more robust. As the annotators are not experts, have the tendency to err in the inferences. To work with this factor, were used in the experiment several annotators. After annotators have reviewed all texts assigned to them, it is necessary to analyze the correlation between different classifications made for the same sentence. We call such activity applying the rule of congruence. As distinct from the proposal by Wilson, who trained the annotators in order to ensure standardization and assertiveness among them, the congruence rule aims to achieve this assertiveness among previously untrained annotators. This rule assumes that in order to reach that goal, it is necessary that the same sentence is annotated by a minimum number of people. For this research we used the minimum number of 3 people. The congruence between annotations of these people ensures standardization and assertiveness desired, in other words, if the 3 annotators have classified a sentence with the same polarity, then this sentence has this polarity. Conversely, the sentences that do not have matching are considered as ambiguous polarity. It is worth noting that the approach of the congruence rule prevents the evaluation of a sentence is skewed with the opinion of one person.

The output generated by this rule is a set of labeled sentences to create training files. These files are used to train new Naive Bayes classifiers, which will be referred as contextualized classifiers and which are able to classify sentences into three polarities: positive, negative and neutral. In particular, the neutral polarity can identify three types of phrases:

1) Subjective phrases with neutral polarity
2) Objective phrases
3) Textual waste that may not have been eliminated by the first process.

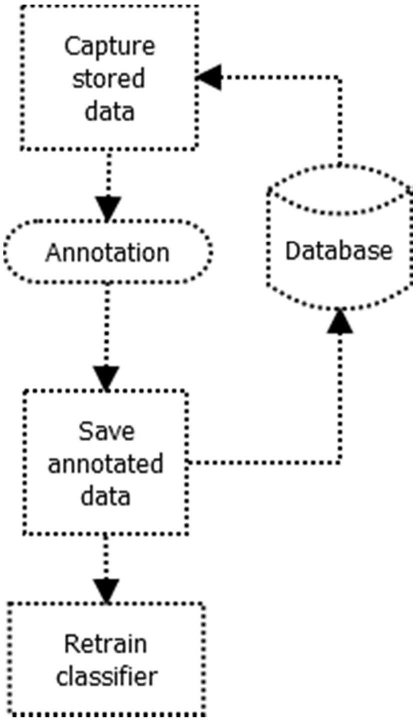

Fig. 2 Scheme for second process of the Sentiment Analyzer

IV. Data Analysis

A. Experiment

The initial data collection fed the Sentiment Analyzer with 180 URLs that contains articles from the entrepreneurship blog. These articles had 4563 sentences, which were classified by the generic classifier. Fig. 3 shows the distribution of positive and negative sentences after inferences have been done. All URLs and articles were stored in web platforms database.

After, people were selected to be annotators in order to analyze the sentences stored in database. As the articles are related to Tech Startups and Business Administration, we selected annotators who had expertise in these areas.

Based on this, all annotators have the following profile:

1) To be Brazilian and to live in Brazil
2) Master reading in English
3) To have experience of at least two years with entrepreneurship or business management

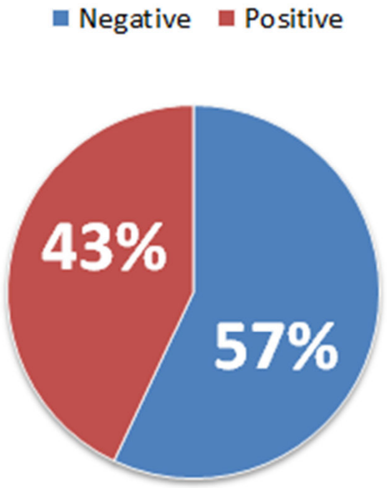

Fig. 3 Distribution of positive and negative sentences after first classification

This profile was intended to ensure that all annotators were inserted in the same context and had the same level of skill with the English language (being neither an expert nor a beginner). Altogether 9 annotators participated in the research, which were divided in two groups, Startup 1 with 6 members and Startup 2 that had 3 members. Each of the groups play a different role in the experiment as will be explained later. It is worth highlighting that the annotators have not received prior training on how performing their tasks within Sentiment Analyzer. To address this factor, it was recommended to all annotators that read the tutorial on the web platform. Moreover, it is very important to mention that they were not trained to identify the polarity of the sentences, because this experiment has the assumption that the annotators do not need to be specialized in this activity to produce relevant data. To group Startup 1 were selected 10 articles to be annotated and to group Startup 2 were selected 5 articles to be annotated. These 15 articles yielded 439 annotated sentences with 309 belonging to the Startup group 1 and 130 related to the group Startup 2. These 15 articles yielded 439 sentences annotated with 309 belonging to the Startup group 1 and 130 related to the group Startup 2. All data (number of annotators and amount of generated sentences annotated) are significant within the context of this paper. Compared to Wilson [27], which used three different annotators that annotated 13 documents with a total of 210 sentences, the amount of data produced in this study was larger, indicating that the sample used is sufficient to evaluate the subsequent stages of this experiment.

Since annotations were made, the next step was to build the training files to generate new classifiers. For this activity the database produced by the Startup group 1 was used. It is worth remembering that on this basis we used the congruence rule cited in section III. Since it was established the first training set, consisting of:

1) A file with 98 positive sentences
2) A file with 37 negative sentences
3) A file with 228 neutral sentences

It is apparent that there is unevenness between the amounts of sentences in each of the files. In order to deal with this situation and study the influence of sentences that belong a given category can impact the outcome of a classifier, were created four different training files. The four files are: file non-balanced, file balanced by filling, file balanced by withdrawal and file balanced by both. The first was created from the union of simple sentences present in all three files mentioned. The second training file was created from the addition of a set of sentences (positive and negative) present in the Cornells dataset until the number of sentences in each category was equal to the number of neutral sentences. The third training file is formed by removal of a few sentences of positive and neutral unbalanced training file so that the number of sentences in each category was equal to the number of negative sentences. Finally, the last training file consists of merging the two previous approaches so that it does not contain the number of sentences arising from generic base greater than the number of sentences coming from the annotators. The classifiers created receive the same name of training files. Importantly, each of the training files created generates classifiers that have different results.

As the classifiers were created, they are tested for accuracy. The data produced by startup group 2 were used to conduct this analysis. Accuracy is defined as the degree of agreement between the inferences made by the annotators and those performed by classifiers relative to unambiguous sentences produced by the group Startup 2. Table I shows the degree of agreement for each of the test articles.

B. Analysis of Results

As stated earlier, the main goal of this experiment is to test the hypothesis that is possible to generate sentiment analysis classifiers that present similar results to that show similar results to those created with input from specialist annotators using data produced by untrained annotators. Based on Table I it's possible to realize that the generic classifier has a very low accuracy on the test articles. This result was expected since the distribution of unigrams in the general context of distribution is different in context, which means that inferences made by the classifier are weak. Moreover, the four new classifiers generated have superior results.

The classifier balanced by filling obtained the worst results among the new ones. Despite having been created from training files with the same amount of sentences, most of the attributes produced from the positive and negative sentences comes from the training files from the base of Cornell University, i.e., are not contextualized to the theme blog entrepreneurship . Given this factor, we can say that when the number of sentences not contextualized is greater than the contextualized sentences, the classifier produced receives negative influence.

The unbalanced classifier showed the best results. We emphasize that this classifier tends to classify sentences as neutral, given the greater number of neutral sentences used for training in relation to the number of positive and negative sentences. The consequence is that were produced more attributes that indicate neutrality in sentences than others. This affects the probability calculations classification of the Naïve Bayes algorithm.

TABLE I
DISTRIBUTION OF POSITIVE AND NEGATIVE SENTENCES AFTER FIRST CLASSIFICATION

Text	Generic Classifier	Non-balanced Classifier	Classifier balanced by filling	Classifier balanced by withdrawal	Classifier balanced by both
Text 1	28.6%	67.9%	32.1%	53.6%	32.1%
Text 2	50%	80%	60%	40%	40%
Text 3	19%	71.4%	19%	66.7%	38.1%
Text 4	6%	73.3%	26.7%	53.3%	26.7%
Text 5	22.7%	77.3%	50%	63.6%	68.9%
Average	25.3%	74%	37.4%	55.4%	41.2%

It is interesting to note that the classifier balanced for both and the classifier balanced by withdrawal obtained intermediate results in relation to the remaining balanced classifiers. This result indicates that, the smaller the interference of not contextual attributes in the training set, the more accurate becomes the classifier within the context of application. In particular, when there is no influence of contextual attributes, the results are more satisfactory, as can be seen from the results of the two best classifiers.

In contrast to the research produced by Wilson [27], this study provides improvements over the model proposed by the researcher. It uses the intersections in analysis of untrained annotators to generate training data analysis for sentiment analysis classifiers. Another factor to be considered is the time of implementation of this approach, as this research used two months to collect the data of the annotators and continued the experiment described here.

V. CONCLUSION

This paper presented a framework to work creating classifiers sentiment analysis more robust and contextualized with minimum effort and time optimized implementation. The results show that non- contextual classifiers have low accuracy when applied in the texts of entrepreneurship blog.

Moreover, we used the intersection between views of different annotators to produce a consistent basis for training Naïve Bayes algorithm. In other words, we used the collective intelligence to produce better results. It is worth highlighting that this strategy avoids creating classifiers biased by subjective view of a person.

All this research was performed within one web platform, Sentiment Analyzer, which is able to perform all procedures described in this study and are necessary for proper implementation of the framework presented here.

The experimental results obtained in this study validate the hypothesis that classifiers sentiment analysis can become more accurate from training with these training files contextualized. All new classifiers generated showed better results than the generic classifier used as the basis of comparison.

REFERENCES

[1] Internet World Stats, Usage and Population Statistics. http://www.internetworldstats.com/, January 2013.

[2] Movie Review Dataset. http://www.cs.cornell.edu/people/pabo/moviereview-data/, January 2013.

[3] Natural Language Toolkit. http://www.nltk.org/, January 2013.

[4] Statsoft - Technical Notes about Naive Bayes Classifier. http://www.statsoft.com/Textbook/Naive-Bayes-Classifier, January 2013.

[5] Wordnet, a Lexical Database for English. http://wordnet.princeton.edu/, January 2013.

[6] Alina Adreevskaia and Sabine Bergler. Mining Wordnet for Fuzzy Sentiment: Sentiment Tag Extraction from Wordnet Glosses. In 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 209–216, 2006.

[7] The Nielsen Company. Global Faces and Networked Places, a Nielsen Report on Social Networkings New Global Footprint. Technical Report, Nielsen Company, March 2009.

[8] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In WWW, pages 519–528, 2003.

[9] Emanuel de Barros Albuquerque Ferreira. Análise de sentimento em redes sociais utilizando influência de palavras. Trabalho de Graduação - Universidade Federal de Pernambuco - UFPE. Departamento de Ciênciada Computação, Dezembro 2010.

[10] Vasileios Hatzivassiloglou and Kathleen McKeown. Predicting the Semantic Orientation of Adjectives. In Philip R. Cohen and Wolfgang Wahlster, Editors, ACL, Pages 174–181. Morgan Kaufmann Publishers / ACL, 1997.

[11] Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In COLING, Pages299–305. Morgan Kaufmann, 2000.

[12] Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM.

[13] Bin Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer, 2008.

[14] Soo min Kim and Eduard Hovy. Automatic Identification of Pro and Con Reasons in Online Reviews. In Proceedings of COLING/ACL Poster Sessions, pages 483–490, 2006.

[15] Subhabrata Mukherjee. Sentiment Analysis - A Literature Survey, June 2012. Indian Institute of Technology, Bombay. Roll No: 10305061.

[16] Kamal Nigam. Using Maximum Entropy for Text Classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61–67, 1999.

[17] Neil OHare, Michael Davy, Adam Bermingham, Paul Ferguson, Praic Sheridan, Cathal Gurrin, and Alan F. Smeaton. Topic-Dependent Sentiment Analysis of Financial Blogs. In Proc. of CIKM Workshop on Topic Sentiment Analysis for Mass Opinion (TSA '09), pages 09–16, November2009.

[18] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135, January2008.

[19] Bo Pang, Lillian Lee, and ShivakumarVaithyanathan. 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, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[20] Anna Stavrianou and Caroline Brun. Opinion and Suggestion Analysis for Expert Recommendations. In Proceedings of the Workshop on Semantic Analysis in Social Media, pages 61–69, Stroudsburg, PA, USA, 2012.Association for Computational Linguistics.

[21] P.D. Turney. Mining the Web for Synonyms: Pmi-ir versus lsa on toefl. In Proceedings of the 12th European Conference on Machine Learning, pages 491–502. Springer-Verlag, 2001.

[22] Peter Turney. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Pages 417–424, 2002.

[23] J. Wiebe. Instructions for Annotating Opinions in Newspaper Articles. Department of Computer Science Technical Report tr-02-101, University of Pittsburgh, 2002.

[24] Janyce Wiebe. Learning Subjective Adjectives from Corpora. In Henry A. Kautz and Bruce W. Porter, editors, AAAI/IAAI, pages 735–740. AAAIPress / The MIT Press, 2000.

[25] JanyceWiebe, Rebecca Bruce, Matthew Bell, Melanie Martin, and Theresa Wilson. A Corpus Study of Evaluative and Speculative Language. In Proceedings of the 2nd ACL SIGdial Workshop on Discourse and Dialogue (SIGdial-2001), pages 186–195, Aalborg, Denmark, 2001.

[26] JanyceWiebe and Claire Cardie. Annotating Expressions of Opinions and Emotions in Language. Language Resources and Evaluation. In Language Resources and Evaluation (formerly Computers and the Humanities), pages 165–210, 2005.

[27] Theresa Wilson. Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of private states. PhD thesis, Intelligent Systems Program, University of Pittsburgh, 2007.