Scientific Research Publishing

Social Networking, 2019, 8, 74-83
http://www.scirp.org/journal/sn
ISSN Online: 2169-3323
ISSN Print: 2169-3285

"If We Only Knew How You Feel"—A Comparative Study of Automated vs. Manual Classification of Opinions of Customers on Digital Media

Huoston Rodrigues Batista, José Carmino Gomes Junior, Marcelo Drudi Miranda, Andréa Martiniano, Renato José Sassi, Marcos Antonio Gaspar

Postgraduate Program in Informatics and Knowledge Management, Nove de Julho University, Sao Paulo, SP, Brazil
Email: huostonrodrigues@gmail.com, josecarmino@uni9.pro.br, mdrudi@gmail.com, andrea.martiniano@gmail.com, sassi@uni9.pro.br, marcos.antonio@uni9.pro.br

How to cite this paper: Batista, H.R., Gomes Jr., J.C., Miranda, M.D., Martiniano, A., R Sassi, R.J. and Gaspar, M.A. (2019) "If We Only Knew How You Feel"—A Comparative Study of Automated vs. Manual Classification of Opinions of Customers on Digital Media. *Social Networking*, **8**, 74-83.
https://doi.org/10.4236/sn.2019.81005

Received: December 28, 2018
Accepted: January 22, 2019
Published: January 25, 2019

Abstract

The Web development has drastically changed the human interaction and communication, leading to an exponential growth of data generated by users in various digital media. This mass of data provides opportunities for understanding people's opinions about products, services, processes, events, political movements, and organizational strategies. In this context, it becomes important for companies to be able to assess customer satisfaction about their products or services. One of the ways to evaluate customer sentiment is the use of Sentiment Analysis, also known as Opinion Mining. This research aims to compare the efficiency of an automatic classifier based on dictionary with the classification by human jurors in a set of comments made by customers in Portuguese language. The data consist of opinions of service users of one of the largest Brazilian online employment agencies. The performance evaluation of the classification models was done using Kappa index and a Confusion Matrix. As the main finding, it is noteworthy that the agreement between the classifier and the human jurors came to moderate, with better performance for the dictionary-based classifier. This result was considered satisfactory, considering that the Sentiment Analysis in Portuguese language is a complex task and demands more research and development.

Keywords

Sentiment Analysis, Opinion Mining, Social Media

1. Introduction

With the rise of Web 2.0 and the popularization of social networks and communication platforms, the number of people expressing opinions about products, services and their experiences tends to increase [1].

This scenario represents an opportunity for companies to extract insights from this mass of unstructured data [2], and at the same time presents a challenge, considering that the mass of data being handled increases exponentially every day, making manual analysis impracticable [3].

In addition, the marketing department of companies can benefit from these insights. Research shows that 81% of (American) users claim to have done online research on a product at least once. Among those consumers of online reviews, between 73% and 87% state that their purchase was significantly influenced by these opinions [4]. Since it is likely that there are thousands of these online product reviews, analysis of this content cannot be performed manually, and should therefore be performed out automatically.

From this scenario results, the development of the research area is known as Sentiment Analysis, also known as Opinion Mining. Many researches on the classification of expressions of sentiment have already been performed [5] [6] [7], vast majority of them contemplating the English language. However, in other languages, and more specifically in Portuguese, there is a lack of studies that explore the Sentiment Analysis, which is reflected, also, in the lack of lexical resources for conducting research.

The main focus of this research is to contribute to the field of Sentiment Analysis in Portuguese language by presenting a comparative study of the efficiency of an automatic dictionary-based classifier with the classification by human jurors in a set of customer's comments extracted from one of the largest online jobs' companies from Brazil.

Among other objectives, this research aims to highlight the difficulty in conducting research in Sentiment Analysis, as well as particularities of the analysis in Portuguese language.

After this brief introductory section the paper is organized as follows: in Section 2, the theoretical background is presented; Section 3 exposes the methodology of the computational experiments; Section 4 presents the analysis and discussion of results and, finally, section 5 is concluded with the final considerations.

2. Theoretical Background

Sentiment Analysis, also known as Opinion Mining, deals with the automatic extraction of opinions or emotions about products, services and experience from content generated by users on digital media. This is an active area of research and has been extensively studied in different application areas [4] [7] [8].

Currently, social media platforms are popular vehicles for studying consumer sentiment on a large scale and within a natural environment, because of the sig-

nificant share of online conversations that express consumers' thoughts, feelings and opinions about products, brands, and their experiences. Considering this scenario, the automated sentiment analysis receives increasing attention from academia and industry [3] and has become one of the main techniques for dealing with large volumes of textual data. Typically, automated sentiment analysis techniques are used to classify any text-based document into predefined categories, reflecting the polarity of the sentiment referred to in the text.

Automated classification of sentiment expressed in social media conversations is a challenge for several reasons. First, identifying opinions and polarity in texts written in natural language requires a deep understanding of the explicit and implicit, syntactic and semantic rules of language, which, in turn, requires a lot of effort on the part of the researcher [2].

In addition, the Sentiment Analysis in unstructured texts, typical of social media, is a challenging task, due to the informal nature of these texts, which commonly include abbreviations, spelling errors, emoticons, emojis and informal syntax, that current methods of Sentiment Analysis do not adequately support [9].

Automated technologies transform these challenges into an opportunity by eliminating the need for costly manual analysis and prone to errors and biases, using computerized procedures to extract insights from textual content.

Automatic methods for dealing with the Sentiment Analysis usually involve lexical-based approaches [10] and machine learning [5], or a combination of both [6].

Both approaches have advantages and disadvantages, but none of them produces perfect results, which is perfectly acceptable, taking into account the limitations already discussed above.

An example of this can be found in Canhoto and Padmanabhan [9] who undertook a comparative study of automated versus manual analysis of social media content. Their results show low levels of agreement between manual and automated classifications, and it proves that, regardless of the automatic method and the advances, there is still much to be done in this area.

In general, the use of lexical-based approaches has proved to be less effective than machine-learning models from training examples [5]. However, the use of lexical-based methods has the advantage of not requiring pre-processing or labeling, which makes it difficult to adopt methods based on machine learning [6].

The choice of which approach to use is crucial as it affects the accuracy of the classification of sentiment and needs to be carefully aligned with the nature of the data being analyzed.

3. Methodology

This research can be defined as exploratory and experimental [11] [12] and was constructed in three phases: 1) selection of comments; 2) application of the model; and 3) evaluation of the classification model performance.

The selection of the comments originates from the data collected from a form filled out by the candidates upon cancellation of the service. The act of the candidate canceling the service does not necessarily imply that his sentiment is considered negative. This is because the service provided by the company is the repositioning in the labor market. This way, when the customer gets a job, the service ends.

The online employment company targeted for this study is one of the largest in the country and has been active in the Brazilian market since 1996. In their databases there are more than six million comments. However, to make this research possible, we considered only the comments in the period from January to June 2013, totaling 594,788 documents. One criterion adopted to refine the population was the withdrawal of documents that contained less than 50 characters, since they might not represent a complete sentence, leaving 215,462 documents. From this population a random selection of 2500 documents was made, which, according to [13], is sufficient as a sample for the set. In order to define the integrity and uniformity of the data, the same fields suggested by Liu [8] were adopted: Code identifier of the comment, Customer identifier code, Date on which the comment was written and Comment written by the customer.

The classification of opinions can be divided into two main approaches: machine learning or lexicons [6]. Approaches based on machine learning have as main disadvantage the need for a mass of data large enough to train and test the model, only then to be able to apply it to the rest of the data. This requires time and effort to construct a database labeled for model training, but usually presents good classification results [5]. Lexical-based classification models, on the other hand, rely on the use of lexical databases with words of opinion to extract the sentiment of the document [8], and do not require any pre-processing or classifier training. Because of their ease of implementation, they tend to be more commonly adopted for sentiment analysis tasks. However, its efficiency is directly linked to the quality of the dictionary used, which can be a limitation for any work done in any language other than English, for which there are dozens of validated and robust dictionaries [10] [14] [15] [16].

For classification purposes, this research used a dictionary-based model [10] [14]. To make this possible, this research adopted one of the most common dictionaries in the Brazilian Portuguese literature known as SentiLex [17]. SentiLex has, in its current version, 7014 lemmas and 82,347 flexed forms.

The next step was to develop software in Python to perform searches for inflected forms in SentiLex and an algorithm for calculating the sentiment scores associated with each comment, described below:

1) The comment is divided into sentences;
2) For each word in the sentence, a search is made on the SentiLex, verifying if there are flexed forms for the same word;
3) If the inflected form is found, the sentiment score will be saved referring to that form;

4) For the word where SentiLex was matched, it is verified if there is a word that represents negation, up to two positions before its occurrence. If the word representing negation is found, multiply the score found by −1;
5) The sum of all the scores found for the sentence is made;
6) Repeat the process for all sentences;
7) The sum of the scores of all the sentences is done, being this the score associated with the comment.

After the calculation of the sentiment scores, the comments were classified as follows:

- Comments with score > 0 were classified as positive;
- Comments with score < 0 were classified as negative;
- Comments with score = 0 were classified as neutral.

After the calculation of the score and the classification, these results are stored in a database.

The evaluation of the performance of the classification models investigated whether their characterization/classification is reliable. For this, it was necessary to characterize the object several times, and to compare the classifier with human jurors. For this, five volunteer jurors were asked to examine a set of 150 comments and classify them as positive, negative or neutral.

The result of the classification by the five jurors was stored in database. Before verifying the agreement of the classifier with the jurors, the agreement between jurors was analyzed.

The evaluation between Jury x Jurors and Jury x Classifier was constructed using the Kappa index, which describes the agreement between two or more judges, or between two classification methods [18] [19] [20] considering besides the percent agreement, the expected agreement by chance. The Confusion Matrix quantifies how many examples of the set would be classified, with the diagonal representing the best classifications [21]. This result allows visualizing not only the global accuracy, but also its behavior [22].

4. Analysis and Discussion of Results

The classifier based on the SentiLex dictionary obtained the results presented in Table 1 and shows that 47.76% of the opinions were classified as positive, followed by neutral and lastly, negative.

One possible explanation for this may be related to the fact that, in the period corresponding to the data analyzed, the company probably obtained a large number of professionals occupying new jobs, which is reflected directly in the amount of positive comments.

The result of the jurors' classification is presented in Table 2, where there is a concordance between jurors 01 and 02 in negative, neutral and positive comments, and disagreement between jurors 03 and 04 in the neutral comments, but with agreement in the positive comments.

The application of the Kappa index in the jury versus jury ranking is pre-

sented in Table 3, and shows strong concordance between the jury 01, 02 and 05, since the obtained agreement was very good. Thus, the Kappa index provided the quantification of concordance among jurors, ranging from 0.660 to 0.839. That is, they indicate agreement with "Good" to "Very Good" variation, which characterizes satisfactory concordance among jurors, thus allowing a comparison with the dictionary-based classifier.

Table 4 shows the intensity of agreement between the human judges and the classifier as "Small", with a Kappa coefficient ranging from 0.243 to 0.333. It is concluded, therefore, that there is no agreement between the two methods.

Table 1. Classification of comments with Dictionary-based Classifier.

Sentiment	Amount	Percentage
Negative	573	22.92
Neutral	733	29.32
Positive	1194	47.76
Total	2500	100.00

Table 2. Classification of comments by Human Judges.

Juror	Negative		Neutral		Positive		Total	
	Quant.	%	Quant.	%	Quant.	%	Quant	%
1	40	26.67	36	24.00	74	49.33	150	100
2	41	27.33	34	22.67	75	50.00	150	100
3	30	20.00	40	26.67	80	53.33	150	100
4	46	30.67	22	14.67	82	54.67	150	100
5	37	24.67	35	23.33	78	52.00	150	100

Table 3. Application of the Kappa index in the jury versus jury ranking.

	Juror 01 × 02		Juror 01 × 03		Juror 01 × 04		Juror 01 × 05	
Agreement among jurors	132	88.00%	127	84.67%	131	87.33%	135	90%
Agreement by chance	56.10	37.40%	57.10	38.04%	58	38.67%	56.70%	37.83%
Kappa	88	---	0.753	---	0.793	---	0.839	
Kappa standard error	0.042	---	0.046	---	0.043	---	0.037	
Confidence interval	95%	0.727 to 0.890	95%	0.663 to 0.842	95%	0.727 to 0.890	95%	0.763 to 0.915
Agreement considered	Very Good		Good		Good		Very Good	
	Juror 02 × 01		Juror 02 × 03		Juror 02 × 04		Juror 02 × 05	
Agreement among jurors	132	88.00%	124	82.67%	130	86.67%	132	88%

Continued

Agreement by chance	56.10	37.40%	57.00	38.18%	58.6	39.04%	56.10%	37.40%
Kappa	88	---	0.72	---	0.781	---	0.808	
Kappa standard error	0.042	---	0.048	---	0.044	---	0.042	
Confidence interval	95%	0.727 to 0.890	95%	0.626 to 0.813	95%	0.695 to 0.867	95%	0.727 to 0.890
Agreement considered	Very Good		Good		Good		Very Good	
	Juror 03 × 01		Juror 03 × 02		Juror 03 × 04		Juror 03 × 05	
Agreement among jurors	127	84.67%	124	82.67%	119	79.33%	122	81.33%
Agreement by chance	57.10	38.04%	57.00	38.18%	58.8	39.20%	58.30	38.89%
Kappa	0.753	---	0.72	---	0.660	---	0.695	---
Kappa standard error	0.046	---	0.048	---	0.050	---	0.050	---
Confidence interval	95%	0.663 to 0.842	95%	0.626 to 0.813	95%	0.562 to 0.758	95%	0.596 to 0.793
Agreement considered	Good		Good		Good		Good	
	Juror 04 × 01		Juror 04 × 02		Juror 04 × 03		Juror 04 × 05	
Agreement among jurors	131	87.33%	130	86.67%	119	79.33%	130	86.67%
Agreement by chance	58	38.67%	58.6	39.04%	58.8	39.20%	59.10	39.41%
Kappa	0.793	---	0.781	---	0.660	---	0.780	---
Kappa standard error	0.043	---	0.044	---	0.050	---	0.044	---
Confidence interval	95%	0.727 to 0.890	95%	695 to 0.867	95%	0.562 to 0.758	95%	0.693 to 0.866
Agreement considered	Good		Good		Good		Good	
	Juror 05 × 01		Juror 05 × 02		Juror 05 × 03		Juror 05 × 04	
Agreement among jurors	135	90%	132	88%	122	81.33%	130	86.67%
Agreement by chance	56.70	37.83%	56.10%	37.40%	58.30	38.89%	59.10	39.41%
Kappa	0.839	---	0.808	---	0.695	---	0.780	---
Kappa standard error	0.037	---	0.042	---	0.050	---	0.044	---
Confidence interval	95%	0.763 to 0.915	95%	0.727 to 0.890	95%	0.596 to 0.793	95%	0.693 to 0.866
Agreement considered	Very Good		Very Good		Good		Good	

Table 4. Application of the Kappa index to Dictionary-based Classifier vs. Jurors.

	Classifier vs. Juror 01		Classifier vs. Juror 02		Classifier vs. Juror 03		Classifier vs. Juror 04		Classifier vs. Juror 05	
Agreement among jurors	78	52.00%	80	53.33%	86	57.33%	87	58%	86	57%
Agreement by chance	54.80	36.56%	53.90	36.60%	56.8	37.87%	55.50	37.01%	55.80	37.19%
Kappa	0.243	---	0.264	---	0.313	---	0.333	---	0.321	---
Kappa standard error	0.059	---	0.059	---	0.061	---	0.056	---	0.059	---
Confidence interval	95%	0.128 to 0.358	95%	0.148 to 0.380	95%	0.194 to 0.433	95%	0.223 to 0.444	95%	0.205 to 0.437
Agreement considered	**Litle**		**Litle**		**Litle**		**Litle**		**Litle**	

5. Final Considerations

The results of non-agreement between the classifier and human jurors denote the need for an analysis of the nature of the comments. It was observed that the dictionary-based classifier encountered classification difficulties. These difficulties arise due to the nature of the dictionary-based model, since, as stated earlier, its effectiveness is directly linked to the construction of the dictionary used in the research. In Portuguese language there is a marked lack of lexical resources for the development of research related to opinion mining, which, in turn, represents an opportunity for future research.

Taking, for example, the difficulty in classifying the word "*vaga*" (meaning "vague" in English), which in the SentiLex dictionary is considered as an adjective and has negative meaning, however, for the domain addressed in this research, the term is considered as noun and, consequently, of neutral meaning. This, in turn, explains part of the lack of agreement between the two methods.

This problem can be solved by using a domain-specific dictionary. This result implies the need to evaluate other methods, or the creation of a specific dictionary for the domain.

According to Pang and Lee [4], the sentiment and the subjectivity are directly connected to the context and, to a certain extent, dependent on the domain. However, the author states that the general notion of positive and negative opinions is quite consistent across domains. This leads to the conclusion that the results found in this research are relevant, taking into account that the Sentiment Analysis in Portuguese language is a complex task and demands more research and development.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Newman, R., Chang, V., Walters, R.J. and Wills, G.B. (2016) Web 2.0—The Past and the Future. *International Journal of Information Management*, **36**, 591-598. https://doi.org/10.1016/j.ijinfomgt.2016.03.010

[2] Cambria, E. (2013) New Avenues in Opinion Mining and Sentiment Analysis. 7.

[3] Chen, H. and Zimbra, D. (2010) AI and Opinion Mining. *IEEE Intelligent Systems*, **25**, 74-80. https://doi.org/10.1109/MIS.2010.75

[4] Pang, B. and Lee, L. (2008) Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, **2**, 1-135. https://doi.org/10.1561/1500000011

[5] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, **10**, 79-86. https://doi.org/10.3115/1118693.1118704

[6] Medhat, W., Hassan, A. and Korashy, H. (2014) Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, **5**, 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

[7] Ravi, K. and Ravi, V. (2015) A Survey on Opinion Mining and Sentiment Analysis: Tasks, Approaches and Applications. *Knowledge-Based Systems*, **89**, 14-46. https://doi.org/10.1016/j.knosys.2015.06.015

[8] Liu, B. (2012) Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, **5**, 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[9] Canhoto, A.I. and Padmanabhan, Y. (2015) We (don't) Know How You Feel—A Comparative Study of Automated vs. Manual Analysis of Social Media Conversations. *ournal of Marketing Management*, **31**, 1141-1157. https://doi.org/10.1080/0267257X.2015.1047466

[10] Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, **37**, 267-307. https://doi.org/10.1162/COLI_a_00049

[11] Hegde, D.S. (2015) Essays on Research Methodology. Springer Berlin Heidelberg, New York, NY. https://doi.org/10.1007/978-81-322-2214-9

[12] de A. Martins, G. and Theóphilo, C.R. (2017) Metodologia da Investigação Científica Para Ciências Sociais Aplicadas, 3ª Edição. Atlas, Brasil.

[13] Tavakoli, H. (2013) A Dictionary of Research Methodology and Statistics in Applied Linguistics. Rahnamā, Tehran.

[14] Avanco, L.V. and das G. V. Nunes, M. (2014) Lexicon-Based Sentiment Analysis for Reviews of Products in Brazilian Portuguese. 2014 *Brazilian Conference on Intelligent Systems*, Sao Paulo, 18-22 October 2014, 277-281. https://doi.org/10.1109/BRACIS.2014.57

[15] Musto, C., Semeraro, G. and Polignano, M. (2014) A Comparison of Lexicon-Based Approaches for Sentiment Analysis of Microblog Posts. Vol. 59.

[16] Chiavetta, F., Lo Bosco, G. and Pilato, G. (2016) A Lexicon-Based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language. 12*th International Conference on Web Information Systems and Technologies*, Roma, 23-25 April 2016, 159-170.

[17] Silva, M., Carvalho, P. and Sarmento, L. (2012) Building a Sentiment Lexicon for Social Judgment Mining. *International Conference on Computational Processing of the Portuguese Language*, Coimbra, 17-20 April 2012, 218-228. https://doi.org/10.1007/978-3-642-28885-2_25

[18] Siegel, S. and Castellan, N. (1988) Nonparametric Statistics for the Behavioral Sciences. 2nd Edition, McGraw-Hill, New York.

[19] Banerjee, M., Capozzoli, M., McSweeney, L. and Sinha, D. (1999) Beyond Kappa: A Review of Interrater Agreement Measures. *Canadian Journal of Statistics*, **27**, 3-23. https://doi.org/10.2307/3315487

[20] Fleiss, J.L., Levin, B. and Paik, M.C. (2003) Statistical Methods for Rates and Proportions. 3rd Edition, Wiley, Hoboken. https://doi.org/10.1002/0471445428

[21] Landis, J.R. and Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**, 159. https://doi.org/10.2307/2529310

[22] Campbell, J.B. (2002) Introduction to Remote Sensing. The Guilford Press, New York.