Nurses and Midwives in the Digital Age
M. Honey et al. (Eds.)

doi:10.3233/SHTI210743

Automatic Polarity Classifier Model of Opinionso About Prostate Cancer on Facebook®

Grace TM DAL SASSO[a,1] and João AS BUENO[b]
[a] *Universidade Federal de Santa Catarina, Brasil*
[b] *Instituto Federal de Santa Catarina, Brasil*

Keywords. Data mining, language processing, natural, machine learning, prostate cancer, social networks

1. Introduction

Disease prevention campaigns, especially those carried out on the Internet, are of paramount importance today because, through them, it is possible to understand, in real-time, at least partially, if the population is well receiving the campaign [1].

This context requires automatic monitoring, which works as an auxiliary tool for managers in decision-making processes. The mission of extracting useful information from unstructured databases (text, speech, etc.), as in social networks, is still a significant challenge. Many researchers in the Natural Language Processing (PLN) subfield have sought the solution to this question called sentiment or opinion analysis [2,3,5]. Therefore, the purpose of this study is to develop and analyze an automatic opinion polarity classifier model applied to the postings on prostate cancer on the Facebook® page named November Blue in 2018.

2. Approach

This study is an innovative technological production of quantitative nature which we developed in five steps. First, we extracted data from the social network to build a dataset in an electronic spreadsheet. Soon after, in the second stage, we preprocessed the data, removing the suffixes of the words, the words repeated or not relevant in the formation of the sentences.

Then a sampling of the data was graded: positive or negative. Three technology experts performed this ordination named training. [4,5]. The fourth step focused on creating and using the machine learning algorithm, called naive Bayes, which analyzed and processed the training that the expert classifiers performed in the previous step, automatically classifying the remaining material available in the database. [3] Subsequently, we studied the model, measured its accuracy and applied methods to assist

[1] Corresponding Author, Grace TM Dal Sasso, Av. Governador Ivo Silveira 177, ap502 Estreito, Florianopólis, SC, Brazil. CEP: 88075005.

in the visualization of the results, thus making changes to adjustments to the source code [4,5].

For the data extraction and dataset creation process, we used the Netlyc platform for its degree of reliability. Then we used the Python programming language because it is free and multiplatform and applied filters to analyze the stopwords of the Portuguese language from the Natural Language Toolkit (NTLK) library. We subsequently performed the human classification to observe the degree of disagreement or complete agreement. The use and increment of the naive Bayes algorithm were necessary actions since this algorithm, and its methods made the automatic classification work more manageable and with good results. First, however, as indicated and with much attention to detail, we performed the previous steps [4,5].

3. Body

The construction of the dataset consisted of searching and extracting the posts and comments related to prostate cancer, from the page called Novembro Azul, on Facebook®. We used the structure Netlytic, which currently provides the services to students and researchers free of charge of social network analysis and data mining tasks [1,2]. In this first step, which was the extraction of data and creating the dataset, a total of 645 posts and comments were obtained from the fan page Blue November. Then, after exporting and pre-processing the data, 261 messages resulted. Then from the library (NLTK), the messages went through an algorithm to remove the stopwords, the FreqDist, which calculates the frequency of words and the stemming, keeping only the radicals of words [2,4].

In the next step with supervised learning, we classified 261 posts. 70% of this classification, that is, 182 messages, served as training for the algorithm. The remaining messages (79), corresponding to 30% of the classification, did for the algorithm to sort them automatically, and then the variety was compared with the sort performed by the experts. Lastly, this algorithm is trained with the training base that corresponds to 70% of the November Blue database, implemented with NaiveBayes.train, which learns from data already sorted by experts and automatically sorts the data. There was better detail through the use of the confusion matrix than the algorithm classified correctly or incorrectly. Of the 61 sentences classified by experts as positive, the method hit 59, that is, 96.72%. In contrast, in the negative sentences, the percentage of hits dropped many, i.e. of the 18 sentences classified by experts as negative, the algorithm only hit 8 (44.44%), much lower than expected. [1,4,5].This result demonstrates which method has a high ability to classify positive phrases correctly and, on the other hand, has great difficulty in categorising negative expressions correctly. The model obtained in the accuracy test the result of 84.8% in the automatic classification of opinions concerning the experts' classification, thus concluding that the implemented model may help health campaign managers in decision-making processes.

4. Conclusions

The main results of the research showed that the use of the study's automatic classification model can contribute as a support tool in the decision making process, especially for the management team or even for the marketing team of the agency that

commands this type of campaign [1], [5]. It is evident from the results that the text preprocessing step is very important for the automatic classification performed by the naive bayes algorithm, even with almost 85% accuracy, which is considered a good result, a better focus on the pre-text step. processing, especially stemming, and better work with stopwords can greatly increase results.

References

[1] Bhaskar J, Sruthi K, Nedungadi P. (2015). Hybrid approach for emotion classification of audio conversation based on text and speech mining. Procedia Computer Science, 46, 635-643. https://doi.org/10.1016/j.procs.2015.02.112

[2] Chen M. (2020). A guide: Text analysis, text analytics & text mining. Available in: <https://towardsdatascience.com/a-guide-text-analysis-text-analytics-text-mining-f62df7b78747> Access in: June 17, 2021.

[3] Antons D, Grünwald E, Cichy P, Salge TO. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities, R&D Management, 50 (3),329-351. https://onlinelibrary.wiley.com/doi/10.1111/radm.12408

[4] Crannell WC, Clark E, Jones C, James TA, Moore J. A pattern-matched Twitter analysis of U.S. cancer-patient sentiments. J. Surg. Res. 206 (2) (2016) 536–542, https://doi.org/10.1016/j.jss.2016.06.050

[5] Farzindar A; Inkpen D. (2015) Natural language processing for social media. Synthesis Lectures on Human Language Technologies.8(2), 1-166. https://doi.org/10.2200/S00659ED1V01Y201508HLT030