Building a sentiment analysis model for libraries: the CSBNO Consortium approach

Anna Maria **Tammaro**[1,], Michele **Tomaiuolo**[2], Monica **Mordonini**[3], Mattia **Pellegrino**[4] and Riccardo **Demicelis** [5]

[1] *University of Parma, Parma, Italy*
[2] *University of Parma, Parma, Italy*
[3] *University of Parma, Parma, Italy*
[4] *University of Parma, Parma, Italy*
[5] *CSBNO Consortium, Milan, Italy*

Abstract

The CSBNO Consortium investigated the libraries communities during the lockdown and at their reopening, to learn about their wishes and expectations from the library. Sentiment analysis could improve the analysis of data integrating the community's perception of the library in services design. The framework and the methodology of the research are described in the three foreseen phases: Selection and loading of training data, Text processing, Creating a model. The research is in its initial phase and three characteristics will be analyzed: Information access, Library space, Affect service. The findings will support CSBNO to promote innovative libraries by actively engaging with participative communities.

Keywords

Sentiment analysis for libraries; User studies; Participatory approach

1. Introduction

The CSBNO (Culture Socialità Biblioteche Network Operativo) Consortium manages 60 libraries in the Milan area and coordinates the transformation of libraries and innovation of services to make them supporting the changing needs of society. CSBNO collaborates with other innovative European libraries gathered in the NEWCOMER[2] project funded by ERASMUS + which intends to promote the vision of the innovative libraries improving the community. The NEWCOMER Project partners intend to promote innovative libraries by actively engaging with users, in a participatory approach. A Manifesto[3] is shared by all Project NEWCOMER partners.

1.1 How to get to know the library community?

At the beginning of the Covid-19 pandemic in Italy, during the first lockdown from March to May 2020, the CSBNO tried to stay connected to libraries community, informing them that libraries, even if closed, continue the service. The greatest difficulty for CSBNO has been to change the service model from face-to-face services to remote services and understanding communities wishes and expectations from the library. For data collection, more than 30.000 telephone calls were made by librarians from the CSBNO Consortium to members classified as active and inactive members.

Proceedings Name, Month XX–XX, YYYY, City, Country

EMAIL: annamaria.tammaro@unipr.it (A. 1); michele.tomaiuolo@unipr.it (A. 2); monica.mordonini@unipr.it (A. 3); mattia.pellegrino@unipr.it (A. 4); riccardo.demicelis@csbno.net (A. 5)
ORCID: 0000-0002-9205-2435 (A. 1); 0000-0002-6030-9435 (A. 2); 0000-0002-5916-9770 (A. 3); 0000-0002-6592-7451 (A. 4)

[2] https://publiclibraries2030.eu/our-projects/newcomer/

[3] https://davidlankes.org/a-manifesto-for-global-librarianship/

Active members are defined as users enrolled in CSBNO libraries starting from 31/12/2017 and still active in using the loan service, selecting those aged between 25 and 65 years. Inactive members are defined as users enrolled in CSBNO libraries starting from 31/12/2017 but no longer active in using the loan service.

Two datasets were collected:
1) telephone replies received from inactive library members;
2) telephone replies received from active members.

The first dataset concerns responses collected during the lockdown from inactive members, but had been active in the past.

The second dataset concerns responses collected during the lockdown of active members who use the library loan.

At the end of the lockdown in May 2020, the libraries of the CSBNO Consortium participated in a national satisfaction survey called "Library for you" on the perception of the library by users. The aim of the national survey was to analyze the satisfaction towards libraries upon reopening. The questionnaire administered soon after the lockdown allowed respondents to answer qualitative questions about the level of service and to leave open comments that provide additional data for understanding community opinions. The answers of communities concerning the CSBNO libraries have been extracted.

The third dataset concerns the responses to the national Satisfaction survey from CSBNO libraries community.

2. Aims and objectives

The CSBNO Consortium intends to build a sentiment analysis model as a tool to explore community expectations and wishes on which to build a participatory approach for service design. The aim of the research is to establish a data mining model to perform sentiment analysis on qualitative comments collected by libraries. The objective is to test a new analytical method to be used to understand the data collected from community and their year-by-year comparison.

The feedback mechanism most used by libraries in Italy is usually the survey collecting data with a questionnaire, such as "The Library for You" survey. However this data collection has the drawback that it is administered to only active users. To overcome this limitation, sentiment analysis, or opinion mining, can use text datasets with data mining programs. As the name suggests, sentiment analysis involves the analysis and identification of positive and negative opinions and emotions within a given text (Wilson et al. 2005). By building such a model today, future library surveys done by the CSBNO Consortium can be analyzed quickly and effectively to provide an accurate assessment of users' overall perception of specific areas of the library.

2.1 Sentiment analysis for libraries

Sentiment analysis for libraries has never been studied in Italy. The international library community has used sentiment analysis in three ways: using social media, using free answer text of questionnaires, and using other corpora.

An experience that is important for this research was carried out by Canadian libraries by collecting the free text responses of the LibQual questionnaire (Moore 2017). The characteristics analyzed by Canadian libraries were:

- **Information control**: access to information, promotion, skills and bibliographic guides;
- **Library space**: approach to physical or digital space;
- **Affect service**: negative and positive sentiment for service.

These three characteristics analyze the feelings for the two fundamental services of the library seen as access to a collection and physical space. An emotional perception that the library in general arouses in users is added.

3. Methodology

To gain comparative appreciation for respondent feedback over time, the comments of the three datasets collected by CSBNO will be analyzed to track their sentiment and the topics they relate to. To gain control over such a significant amount of data, computer-aided data mining tools will be used to conduct sentiment analysis on the comments of each dataset of the survey. The framework for the sentiment analysis model essentially involves three steps: selection of training data, text processing, creating a model. Two students have been involved in the project.

3.1 Selection and loading of training data

The pre-tagged training data is selected and loaded into the program. To create a template, both text elements and any corresponding sentiment assignments must be selected.

3.2 Text processing

The text considered by the CSBNO Consortium is in Italian. Text preprocessing eliminated minor language differences, such as lowercase versus uppercase letters, pluralization, and tenses, using common stemming and stop-words techniques, to create an accurate text analysis model. However, since some models use the grammatical structure of text, the original plain text is also kept in the dataset, for possible use in the following steps of analysis. Once finished, the training data corpus is used to create positive, negative and neutral vectors of features, to capture the polarized elements that characterize the text of the comment. Those vectors of features are saved for future use.

3.3 Creating a model

Using these vectors of features, the program uses a classification algorithm to create a pattern to separate other unseen comments into positive, negative, or neutral, for sentiment analysis. As an orthogonal task, the selected comments will also be classified according to their topic. This further classification will provide a deeper and more complete understanding of the collected opinions. To verify the accuracy of the models, they are tested on some pre-tagged test data, to measure the precision, recall and accuracy of the classification. The model is saved for future use.

In simplified terms, the most traditional approaches of sentiment analysis work by providing the algorithm with a so-called “bag of words”, that allows it to recognize the words and the groups of words that humans use to express positive and negative opinions. The process is a form of supervised machine learning; pre-tagged datasets are used as training examples to “teach” the computer and create the basis for the classification of future unlabeled information. By providing pre-tagged “positive” (good, polite, excellent, etc.) and “negative” (terrible, shoddy, rude, etc.) words, the data mining software can establish a model that will be applied to future comments to decipher their polarity or whether they have a positive or negative feeling. With the same approach, it is also possible to classify a text according to its specific topic, in a task of topic detection. Alternatively, it is possible to use clustering algorithms to group together texts with similar features, in a non supervised scenario.

In this work, the most representative and consolidated techniques of sentiment analysis and topic detection will be compared. In particular, the best algorithms of different families will be considered, including those based on some notion of geometric distance between samples (i.e. knn, svm) (Huq et al. 2017), decision trees (rf, xgboost) (Hama et al. 2021), probability and statistics (NB) (Laksono et al. 2019), perceptrons and small neural networks (Akhtar et al. 2017). In fact, those algorithms, or their composition (Angiani et al. 2016), have proven their good accuracy over many different datasets, of small and medium size (Tomaiuolo et al. 2020).

However, some newer algorithms have improved the accuracy over larger datasets, exploiting so-called deep neural network architectures, together with more advanced techniques for collecting the vectors of features of the training set. In fact, the traditional vectorization, based on the bag of words algorithm, creates a dataset with a very large number of features and requires an accurate and sensible phase of feature selection, for obtaining the best results. Instead, techniques of word embedding and dense representations (Yadav et al. 2020) are able to map each word in a multidimensional space, where semantically related words are represented as points at a short distance. The vector representing each sample is calculated on the basis of positions of words in this multidimensional space. Moreover, deep neural networks have shown some impressive results in many applications, including sentiment analysis (BERT) (Sun et al. 2019). But these networks are characterized by a very large number of parameters which have to be learned, requiring the use of samples in the order of magnitude of Big Data.

In the present work, these new techniques will also be used, exploiting pre-trained models and additional phases of transfer learning and fine tuning, for adapting the models to the particular task at hand. The additional steps usually require much smaller datasets, than those used to train the whole model.

The research is in its initial step. We intend to analyze a training set of some comments, randomly selected from the responses of the three data sets collected. This training set will be manually reviewed by the two students and labeled as having a positive or negative feeling.

Using the data mining platform these training sets of comments will provide the framework for creating data-specific positive and negative word vectors to power the sentiment analysis model. It is thought to create an additional process to isolate individual topics within the larger comments, allowing for more nuanced sentiment analysis.

4. Conclusions

The sentiment analysis model provides a complementary tool for analyzing quantitative and qualitative results of simple satisfaction survey of active and inactive users for library services. Sentiment analysis application, could facilitate the realization of a participatory approach with communities, allowing a simple and efficient year-by-year analysis of open comments. The CSBNO Consortium expects the sentiment analysis process to provide the means to isolate specific topics based on specified keywords, allowing individual institutions to tailor results for more in-depth analysis.

5. References

Akhtar, M. S., Kumar, A., Ghosal, D., Ekbal, A., & Bhattacharyya, P. “A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis”, in Proceedings of the 2017 conference on empirical methods in natural language processing (2017),, pp. 540-546.

Angiani, G., Cagnoni, S., Chuzhikova, N., Fornacciari, P., Mordonini, M., & Tomaiuolo, M., “Flat and hierarchical classifiers for detecting emotion in tweets”, in Conference of the Italian Association for Artificial Intelligence (2016), pp. 51-64.

Hama Aziz, R. H., & Dimililer, N., “SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier”, in Journal of the Chinese Institute of Engineers (2021), 44(6), pp. 562-572.

Huq, M. R., Ali, A., & Rahman, A., “Sentiment analysis on Twitter data using KNN and SVM”, in International Journal of Advanced Computer Science and Applications (2017), 8(6), pp. 19-25.

Laksono, R. A., Sungkono, K. R., Sarno, R., & Wahyuni, C. S., “Sentiment analysis of restaurant customer reviews on tripadvisor using naïve bayes”, in IEEE 2019 12th International Conference on Information & Communication Technology and System (2019), pp. 49-54.

Moore, M.T., "Constructing a sentiment analysis model for LibQUAL+ comments", Performance Measurement and Metrics (2017), Vol. 18 No. 1, pp. 78-87. https://doi.org/10.1108/PMM-07-2016-0031

Sun, C., Huang, L., & Qiu, X., “Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentences” (2019), arXiv preprint arXiv:1903.09588.

Tomaiuolo, M., Lombardo, G., Mordonini, M., Cagnoni, S., & Poggi, A., “A survey on troll detection”, in Future Internet (2020), 12(2), p. 31.

Wilson, T., Wiebe, J. and Hoffmann, P., "Recognizing contextual polarity in phrase-level sentiment analysis", in Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics, (2005) pp. 347-354.

Yadav, Ashima, and Dinesh Kumar Vishwakarma, "Sentiment analysis using deep learning architectures: a review.", in Artificial Intelligence Review (2020), 53.6, pp. 4335-4385.