Data Mining and Opinion Mining: A Tool in Educational Context

Myriam Peñafiel
DICC
Escuela Politécnica Nacional
Quito, Ecuador
593 999028925
myriam.penafiel@epn.edu.ec

Stefanie Vásquez
DESODEH
Escuela Politécnica Nacional
Quito, Ecuador
593 998376905
maria.vasquez@epn.edu.ec

Diego Vásquez
UGT-Departamento de Ciencias
ESPE
Quito, Ecuador
593 999240609
ddvasquez@espe.edu.ec

Juan Zaldumbide
ESFOT
Escuela Politécnica Nacional
Quito, Ecuador
593 996192500
juan.zaldumbide@epn.edu.ec

Sergio Luján-Mora
DLSI
University of Alicante
Alicante, Spain
034 965903400
sergio.lujan@ua.es

ABSTRACT

The use of the web as a universal communication platform generates large volumes of data (Big data), which in many cases, need to be processed so that they can become useful knowledge in face of the sceptics who have doubts about the credibility of such information. The use of web data that comes from educational contexts needs to be addressed, since that large amount of unstructured information is not being valued, losing valuable information that can be used. To solve this problem, we propose the use of data mining techniques such as sentiment analysis to validate the information that comes from the educational platforms. The objective of this research is to propose a methodology that allows the user to apply sentiment analysis in a simple way, because although some researchers have done it, very few do with data in the educational context. The results obtained prove that the proposal can be used in similar cases.

CCS Concepts

Information systems → Information integration; Computing methodologies → Learning settings; Applied computing → Learning management systems; Applied computing → E-learning.

Keywords

Data mining; Opinion mining; Text mining, Method; Sentiment analysis; Educational data

1. INTRODUCTION

The spread of data from the web that comes with digitization, affordable technology, consumer hardware, social networks, communities, online media, cloud computing, etc., are known as “massive data” or Big data [1]. Every minute there are two million visits to YouTube, 1.7 million Facebook messages posted, and hundreds of thousands of tweets online per minute [2]. This gives us the dimension of the data generated in the web.

By 2018, it is estimated that there will be around 2.67 billion social media users worldwide, compared to 2.34 billion in 2016, affecting work, politics, political deliberation and patterns of communications between other fields through social media [3].

In addition, as individuals, we wander the web, leaving personal marks with regard to who we are, where we are, what we do and what we are interested in - personal information that can be used and marketed due to our ignorance.

Social media is defined as a group of internet-based applications that use the ideological and technological foundations of web 2.0 as seen in Figure 1, and that allows the creation and exchange of user-generated content [4]. Also, social media users are those who said “yes” to “Have you ever used a social networking site like Facebook, Twitter or LinkedIn?” [5]. Therefore, the data obtained from these media constitute an essential input that must be used, and many companies are already doing so [6]. Other important information is the data from the educational context such as learning platforms, chats, and social networks.

Figure 1. Social Media Landscape [5]

All these data that come from the web through all their means, need to be processed to become information. This is done through the use of new tools, methods and techniques such as data mining,

ICoMS 2018, July 15–17, 2018, Porto, Portugal.

ACM ISBN 978-1-4503-6538-3/18/07…$15.00
DOI: https://doi.org/10.1145/3274250.3274263

to obtain knowledge from all types of structured or unstructured data, in a precise, timely and clear way [7]. Opinion mining [8-10] is an emerging data mining technique that applies artificial intelligence at different levels for the processing of texts, which contain opinions, using data processing to classify the opinion found into a positive, negative or neutral feeling, obtaining values of polarity between– 1 and 1.

To exemplify how this works, we can use a study that applied sentiment analysis to the data of the learning platform of the National Polytechnic School, Higher Education Institution in South America taken as a case study.

This paper is an application of data mining using computational techniques such as text mining and analysis of sentiments with the objective of evaluate the open questions of the online surveys conducted by the teachers of the university. The results have allowed obtain relevant information in relation to the time that the teachers use when incorporating online platforms in the process of teaching learning, acceptance or rejection of the use of these tools by teachers, etc.

2. DATA MINING

Data mining is defined as the extraction of implicit, previously unknown, and potentially useful information from data. Data mining is used in a wide variety of fields including business, bioinformatics, military, education, communication, web mining, image processing, diagnostics, marketing, sales and other applications [11].

Data mining is used to discover hidden patterns and relationships in large volumes of data. The information discovered helps to make more effective decisions [11].

According to [12], the five current research areas with regard to data mining are web data mining (Web mining), mining of large volumes of data (Big data), mining of data flows (Data stream mining), mining of educational data (Educational data mining), data mining of health care (Healthcare data mining).

2.1 The Process of Data Mining

The process of data mining, from the manipulation of data to knowledge, goes through different stages [13]:

- Data collection: this consists of obtaining data from its sources.
- Pre-processing: the data is converted into an appropriate format (modified data).
- Mining of data: the data are processed using already established and new techniques for the exploitation of hidden knowledge.
- Interpretation of the results: the results are visualized and summarized to simplify the decision-making process.

Data mining techniques can be classified according to the algorithms that are used as supervised or predictive and unsupervised or knowledge discovery. Supervised algorithms predict the value of an unknown attribute based on the relationship with the known attributes, performing a training and testing process. In contrast, unsupervised or knowledge discovery algorithms discover patterns and trends in current data, so they are not used for predictions.

2.2 Text Mining

The main objective of text mining is to extract information from semi-structured or unstructured text using text mining techniques, with or without supervision [14]. Text mining uses the processing of natural language and automatic learning (machine learning) [6] in order to classify the data to obtain patterns or models that generate knowledge.

3. OPINION MINING

Opinion mining [8] is an area that has currently aroused the interest of many investigators in recent years.

The purpose of the opinion mining is to determine the user's attitude in the analyzed text through research, analysis and the extraction of subjective texts that imply the opinions of users, their preferences and their sentiments [1].

This field of research is multidisciplinary in nature, since it involves other fields of research such as natural language processing, computational linguistics, information retrieval, machine learning, and artificial intelligence among others [15]. In terms of the techniques of data mining applied to the opinion mining [14], the most commonly used are Naive Bayes and Support Vector Machine [16].

For opinion mining, it is recommended that users carry out the following phases: creating a framework for the construction of the lexicon, and the extraction of the characteristics [1], [8].

3.1 Construction of the Lexicon

With regard to the framework of the construction of the lexicon, human language develops two types of information: objective and subjective. The line of research associated with subjective facts is related to the opinion mining, whose critical information is related to the interpretation of human subjective feelings. The construction of the lexicon or corpus is usually a prerequisite for the opinion mining. The opinion mining in terms of data in the Spanish language becomes a major obstacle, because the corpus in this language is limited [17]. Consequently, there are only a few studies that have analyzed text in Spanish [18].

3.2 Extraction of the Characteristics

The extraction process consists of labeling the texts in the corpus as being either positive or negative. The method is general and can be applied to any text collection. The characteristics can be themes, objects and relationships. The extraction of themes is where the subject and the feeling of a phrase is analyzed. The extraction of objects is one of the fundamental tasks of the opinion mining, the aim of which is to extract the object from a review of phrases, in order to identify all the characteristics of events in a given text.

The extraction of relationships are relational representations to facilitate document classification. In addition, it solves the problem of the correlation and omits duplicates.

Opinion mining can be used to obtain the positive or negative comments with regard to any product or service found on the web. For example, if we take one of the social networks of the web pages linked to government services, we can see comments made by the readers. Such comments act as the source for obtaining the degree of polarity by applying the opinion mining.

Evaluating the opinion of the users in a positive, negative or neutral sense, based on a polarity value that goes from -1 to 1, is the product that is obtained from applying opinion mining, allowing the user to find trends in the data.

4. METHODOLOGY TSA (TEXT ANALYSIS- SENTIMENT ANALYSIS)

4.1 Phase 1 - Text Analysis

Text analysis consists of the distinction of the different elements of a body of text, and consists of the following activities as you can see in the Figure 2:

"**Step 1: data collection".** This is the grouping from different sources, text files, spreadsheets, XML, JSON or any structured or semi-structured data source.

"Step 2: data cleaning". This consists of data cleansing, or the elimination of noise. This is based on the use of two techniques of lemmatization and tokenization.

The first technique uses lemmatization: this consists of counting, as a single element, words that share the same motto or root of a word. For example, the words computer, computational, computers, have the same root or motto - "computer". Consequently, they are counted as a single element.

The second technique is tokenization. This is a method of lexical analysis and consists of a normal count of the words that make up a phrase. For example, the phrase: "Computers respond better to user commands", would have the following elements:

1: <The> 2: <computers> 3: <answer> 4: <best> 5: <a> 6: <the> 7: <commands> 8: <del> 9: <user>

In addition, in this stage, the most representative elements are counted, such as long words. Words of less than four characters are discriminated in such a way that when combining the aforementioned techniques, we would have the following elements:

1: <compute *> 2: <answer> 3: <best> 4: <commands> 5: <user>

"Step 3: frequency of words". This results in the counting of the elements omitted in step 2**"**. This depends on the technique used. The result of this phase is the frequency of refined words. This technique consists of a normal count of the most frequently repeated words in such a way that we would have, in the example mentioned in the previous step, the most repeated word **"**compute *****"**, in that it is repeated twice. By applying these techniques to a larger dataset, you can see the utility of it.

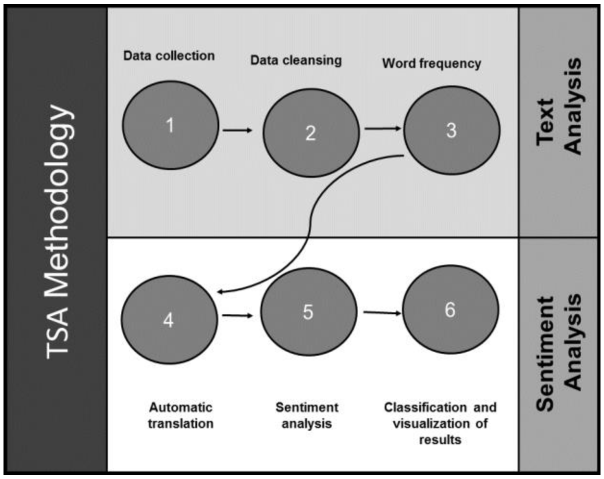

Figure 2. Methodology of text and sentiment analysis (TSA)

4.2 Phase 2 – Sentiment Analysis

Sentiment analysis or Opinion mining is based on determining the polarity of a sentence, whose values are expressed in the range of -1 to 1 in relation to the negativity, neutrality or positivity of the text analyzed. In this phase the following steps are carried out:

"Step 4: automatic translation". This uses automatic translation tools which allows communication with the Google Translator application programming interface to translate the text into the English language.

A rigorous translation process carried out by professional translators would be desirable and would guarantee the results of the proposed proposal. However, data mining processes are applied to large volumes of data, where manual translation would not be possible, or would be very expensive.

The translation is applied to the original sentence without any modification. Depending on the volume of the data, all the data can be translated, or a sample of the words with the highest frequency obtained in Phase 1 can be obtained. The translation process is then applied only to the sentences (opinions) containing those frequently mentioned words. If the investigation is in the English language, step 4 should not be carried out.

"Step 5: Sentiment analysis or opinion mining". The sentences translated in the previous step are used and sentiment analysis is applied properly.

The opinions of the users can be of different lengths. For example: "We cannot have an open door to migrants from the European Union between now and the end of the Brexit process. It is time to control the issue of immigration". Due to the nature of the sentence, each sentence must be subdivided into sentences, and the feelings of each of these sentences analyzed. The TextBlob libraries allow us to perform this task.

For this it is recommended that the user works with the Python libraries TextBlob and vaderSentiment, which give them the result of the processed polarity values. For example, for the phrase "This process has only brought us difficulties and we want it to end", the script returns the polarity value of -0.6249, which means that it is a negative feeling, that is, a complaint or expressing discomfort.

Opinion mining is based on a comparison between a set of training data and a set of test data. TextBlob and vaderSentiment have their own training set that has been used before in other studies [19].

In addition, TextBlob and the training sets of vaderSentiment are very complete libraries that were trained using famous books, movie scripts, social networks, and blogs among others.

"Step 6: classification and visualization of results". From the sentiment analysis of step 5, we obtain a file that contains the data processed with the results obtained, whose information corresponds to the final result of applying the methodology.

5. CASE STUDY

Here is one of the open questions that could not be considered in previous studies due to its complexity of evaluation[21-22]. The data considered for this case study correspond to the surveys conducted at the university in the period from September 2014 to January 2015, applied to teachers to make a diagnosis of the perception of the use of virtual classrooms as a teaching tool in the classroom.

In the research 177 online surveys were collected, representing 47.2% of the university's full-time faculty. We obtained 54 responses to the open question selected for this case study, which corresponds to 30.5% of the total of surveys. The question evaluated was:

Question 28: We invite you to give us your suggestions and recommendations.

Once applied the TSA methodology, here we will only describe the results obtained with the experiments carried out and the discussion of them.

5.1 Phase 1- Text Analysis

The highest frequency words from this case study can be seen in Figure 3, it should be emphasized that because the number of data was limited, the words with the lowest frequency were not eliminated, how as it should be done when the amount of data is high as proposed by the methodology.

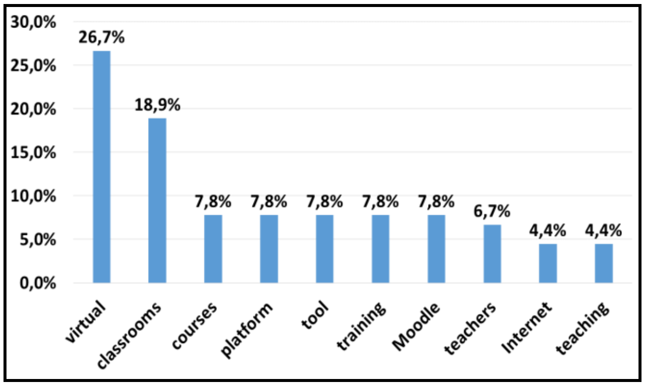

Figure 3. Words with highest frequency result

The results obtained in Phase 1 of this case study reflect the words that arouse the greatest interest of teachers because the open question asked for the suggestions and recommendations, i.e., the concerns of teachers. Now in the Phase 2, we want to know if the interest in these words expresses a positive, negative or neutral feeling on the part of those involved.

5.2 Phase 2- Sentiment Analysis

In the Phase 2 of the methodology to validate the process of sentiment analysis, two methods TextBlob and vaderSentiment, were applied to guarantee the process and to analyze the results, Figure 4 reflects the two scenarios.

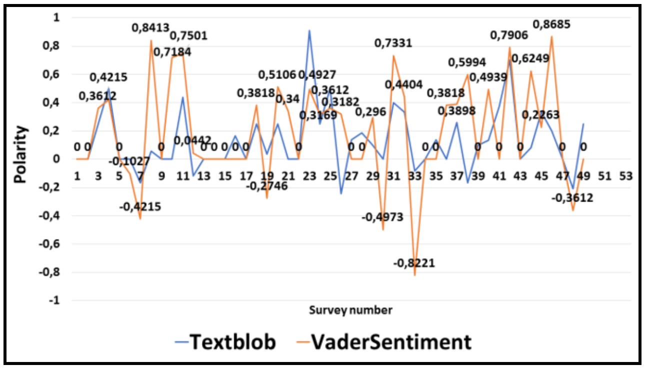

Figure 4. Values of polarity with two methods

By analyzing the results of Figure 4, it can be seen that there is a high correlation with the values obtained by applying the two methods, in the values that approach the polarity extremes. Also, although there are many neutral values, as mentioned [20], they help to improve the overall accuracy of the classification.

Since the difference in polarity values obtained by applying the two methods is not significant, as shown in Table 1. In order to improve the accuracy of the results it is opted to lead to a value equal to the calculation of the average between the two values of polarity obtained from the sentiment analysis applying the two methods.

Table 1. Difference between values of sentiment analysis applying the two methods

Sentence	Text Blobs x1	Vader Sentiment x2	|x2-x1|
Better and more extensive training is required for the use of Moodle.	0.248	0.317	0.069
The time dedicated to the planning and structuring of the contents and activities.	0.047	0.153	0.107

In order to interpret the polarity results, the data were grouped into polarity ranges by applying the Galton-Mac Law to calculate the number of classes of polarity values as a function of the number of data. For this case study with 54 responses seven ranges were obtained as shown in Table 2.

Table 2. Values of Sentiment Analysis

Class	Lower polarity	Upper polarity	Percentage by group
1	-0.45272	-0.28157	5.66%
2	-0.28157	-0.11043	3.77%
3	-0.,11043	0.06072	32.08%
4	0.06072	0.23186	20.75%
5	0.23186	0.40301	20.75%
6	0.40301	0.57415	9.43%
7	0.57415	0.74530	7.55%

5.3 Discussion of Results

Table 2 shows the results obtained in case study and Phase 2. It can be concluded that the suggestions and recommendations are given by the teachers related to the words that obtain the highest frequency. The Figure 4 reflects a significant value of neutral and positive sentiment with a value of 90.57%, with only a 9.43% negative feeling in classes 1 and 2. This information is relevant to be considered as positive points of attention in the use of this educational tool in the classroom. One of the major concerns of teachers is the virtual classroom; it would be necessary to consider the elements derived from it, such as training in the tool, instructional design, to improve the use the virtual classroom as a tool.

6. CONCLUSIONS AND FUTURE WORK

The proposed methodology can be used to evaluate open or opinion questions in the educational context, as support for traditional evaluation methods.

The information obtained can be used to make decisions that should help improve the teaching/learning process with the information obtain of the learning platforms.

As future work, it is proposed to continue experimenting with the TSA methodology to evaluate open or opinion questions in research in the field of social networks, in order to refine it and obtain more precise results. The automation of the methodological proposal would also be very interesting.

7. ACKNOWLEDGMENT

Our thanks to VIPS of Escuela Politécnica Nacional in the project PII-03-16.

8. REFERENCES

[1] Nanli, Z., Ping Z., Weiguo, L. and Meng, C. 2012. Sentiment analysis: A literature review. In *International Symposium on Management of Technology* (ISMOT), 572-576, DOI= https://doi.org/10.1109/ISMOT.2012.6679538.

[2] Warren-Payne, A. 2016. 13 epic stats and facts from The State of Social webinar. https://www.clickz.com/13-epic-stats-and-facts-from-the-state-of-social-webinar/110510/2016

[3] Statista. 2017. The Statistics Portal, https://goo.gl/6fdPTu

[4] Kaplan, A. and Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of Social Media, *Business horizons,* 53, 1, 59-68. DOI= https://doi.org/10.1016/j.bushor.2009.09.003.

[5] Cavazza, F. 2018. Social Media Landscape. https://fredcavazza.net/G.

[6] You, Q. 2016. Sentiment and Emotion Analysis for Social Multimedia: Methodologies and Applications. In *Proceedings of the 2016 ACM on Multimedia Conference* (MM '16). ACM, New York, NY, USA, 1445-1449. DOI= https://doi.org/10.1145/2964284.2971475.

[7] Gandomi, A. and Haider M. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35, 2, 137-144.

[8] Liu, B. 2012. *Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies*. San Rafael: Morgan & Claypool Publishers.

[9] Del-Vicario, M., Zolloa, F., Caldarellia, G., Scalab, A. and Quattrociocchia, W. 2017.Mapping social dynamics on Facebook: The Brexit debate. *Social Networks*, 50, 6-16.

[10] Immigration, Refugees and Citizenship Canada (IRCC). 2017. https://twitter.com/CitImmCanada

[11] Miller, L. D., Soh, L. K., Samal, A., Kupzyk K. and Nugent, G. A. 2015. Comparison of Educational Statistics and Data Mining Approaches to Identify Characteristics that Impact Online Learning. *JEDM-Journal of Educational Data Mining*, 7, 3, 117-150, DOI= https://doi.org/10.1016/j.socnet.2017.02.002

[12] Almasoud, A. M., Al-Khalifa, H. S. and Al-Salman, A. 2015. Recent developments in data mining applications and techniques. In *2015 Tenth International Conference on Digital Information Management* (ICDIM), 36-42.

[13] Zaldumbide, J. and Sinnott, R. O. 2015. Identification and Validation of Real-Time Health Events through Social Media. In 2*015 IEEE International Conference on Data Science and Data Intensive Systems,* 9-16, DOI= https://doi.org/10.1109/DSDIS.2015.27.

[14] Cai, K., Spangler, S., Chen, Z and Zhang, L. 2008. Leveraging Sentiment Analysis for Topic Detection. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* - Volume 01 (WI-IAT '08), 1. IEEE Computer Society, Washington, DC, USA, 265-271. DOI= http://dx.doi.org/10.1109/WIIAT.2008.188.

[15] Feldman, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM* 56, 4 (April 2013), 82-89. DOI= https://doi.org/10.1145/2436256.2436274.

[16] Souza, E., Vitório, D., Castro, D., Oliveira, A. and Gusmão, C. 2016. Characterizing Opinion Mining: A Systematic Mapping Study of the Portuguese Language. In *International Conference on Computational Processing of the Portuguese Language* (PROPOR), 122-127.

[17] Liu, Y., Yu, X., Chen, Z. and Liu. B 2013. Sentiment analysis of sentences with modalities. In *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing* (UnstructureNLP '13). ACM, New York, NY, USA, 39-44. DOI= https://doi.org/10.1145/2513549.2513556.

[18] Martínez-Cámara, E., Martín-Valdivia, M.T., Ureña López, L. A. and Mitkov, R. 2015. Polarity classification for Spanish tweets using the COST corpus, *Journal of Information Science*, 41, 3, 263–272.

[19] Hutto, J. and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text, In *Eighth International AAAI Conference on Weblogs and Social Media.* DOI= https://doi.org/10.1109/TLA.2016.7437226.

[20] Estevao da Silva, L. A. 2016. A Data Mining Approach for Standardization of Collectors Names in Herbarium Database. *IEEE Latin America Transactions*, 14, 2, 805-810. DOI= https://doi.org/10.1109/TLA.2016.7437226

[21] Peñafiel, M., Lujan-Mora, S., Vintimilla, LM. and Pozo P. 2015. Analysis of the usage of virtual classrooms in the National Polytechnic School of Ecuador: Teachers' perception, In *Information Technology Based Higher Education and Training* (ITHET), Lisbon, PT, 1–6. DOI= https://doi.org/10.1109/ITHET.2015.7218015.

[22] Peñafiel, M., Vásquez, S. and Lujan-Mora, S. 2016. Use of Virtual Classroom: Summarized Opinion of the Stakeholders in the Learning-Teaching Process, In *Information Technology Based Higher Education and Training* (CSEDU), Rome, IT, 314–320. DOI= https://doi.org/10.5220/0005797603140320