

Feelings Detection System – a Proposal

Arminda Guerra Lopes¹, Ana Dias², Joana Salgueiro², Eurico Lopes³

¹Instituto Politécnico de Castelo

Branco/EST & Madeira Interactive Technologies Institute, Portugal
aguerralopes@gmail.com

²Instituto Politécnico de Castelo

Branco/EST, Portugal

anocas.margarida@gmail.com, joanasalgueiro97@hotmail.com

³Instituto Politécnico de Castelo

Branco/EST & QRural – Qualidade de Vida no Mundo Rural, Portugal

eurico@ipcb.pt

Abstract. The feelings detection is the study of the human feelings, attitudes and opinions and their classification about a specific topic. This paper consists of carry out the feelings (rever sentido da frase) analysis of a population, expressed in written comments from online newspapers about several subjects. The motivation for this project results from the local newspaper agents who suggested the development of this application for their use. After looking into several algorithms able of doing this analysis it was decided to use the algorithm SentiStrength, which allows words classification based on the punctuations stored in its dictionary as positives or negatives. “To feed” the algorithm authors used comments from online newspapers through a web spider, storing it in a text file that will be the algorithm input. The results will be achieved through the Media reach by the SentiStrength analysis in which the positive and negative feelings are expressed and presented in the extracted comments.

Keywords: SentiStrength, sentiments, detection of sentiments, digital inclusion

1 Introduction

Technological change invades most areas of society and many different aspects of our lives. The utilization of technologies, such as the Internet, increased across all sectors of society. Digital devices and applications play key roles in our daily life and in a wider range of the society. This gives us insight to develop a system, which can contribute to the inclusion of different people in digital world.

The proposed system is an example of public participation in society. Three local newspapers suggested the incentive. The goal was to implicate public participation through comments of newspapers’ news about the written texts in order to get more ideas about how to address the narratives.

Conversely, they wanted to get data about a subject and then, inform others about people’s opinions.

The main purpose of this paper is to present the analyses of the principal techniques of feelings detection in a population, to verify the viability of the development of a system prototype that allows to withdraw people's opinions about different themes. Comments are extracted from online newspapers expressing feelings related to the news. During this work's development some issues emerged, such as, the way to extract newspapers comments and how it would be analyzed. As a solution to our problems we used a Web Spider [14], which allows access to the intended information and analysis of the extracted comments. An already existent algorithm was used, the SentiStrength [6]. The algorithm permits to evaluate words as emoticons (faces that show feelings) classifying each word or emoticon.

In summary, authors tried to focus on: the meaning of the detected feelings from a population; the use of a common algorithm for feelings detection; the solution to access the information from the newspaper; and the whole process from information gathering to its analysis. These steps permitted to test the project proposal viability. Finally, we intended to define the requirements to develop the prototype tool for the population feelings' analysis of a specific event and time.

The final output has several components, expressed in text form, related to a subject. This work will ensure that the users opinion about certain news, expressed on online platforms newspapers, will be understood as well as their feelings about the expressed comments.

2 Related Work

This section presents the background information about sentiments analysis studies and the algorithms, generally, used for this subject.

2.1 Sentiment Analysis

Sentiment analysis deals with the computational treatment of opinion, sentiment and subjectivity in text. Throughout the use of technologies people's opinions can influence in shaping the opinions of others. There are hundreds of papers published on the subject: [2], [7], [8], [9], [15]. The factors behind this include: the rise of machine learning methods in natural language processing and information retrieval and the availability of datasets for machine learning algorithms, among others.

2.2 Tools for Sentiment Classification

The feelings analysis as referred in [12] is the classification of human feelings, attitudes and opinions about a topic expressed in text or speech.

Despite not being a very discussed topic, there are some tools that can help in the sentiments classification expressed in texts or words. In this research some algorithms examples are presented:

- **Emoticons:** as referred in [1] is a tool that analysis emoticons and allows to withdraw the feelings itself. Concerning its efficiency, it is more efficient when used with another algorithm because it only refers to the feeling expressed and not the value.
- **LIWC:** this program searches words belonging to the text we want to analyze and verifies if these words exist in its dictionary. If these words exist, they will be stored in the right category. The words are analyzed separated by categories, for instance, articles, personal pronouns, positive and negative feelings [3].
- **SentiWordNet:** according to the Princeton University [13] this algorithm is based on a wordnet dictionary that is compounded by a wide range of nouns, verbs, adjectives and adverbs, which are grouped in sets of synonyms called synsets to express a different concept. The results from this method are divided into three categories as positivity, negativity and objectivity. Each one is related with a color.
- **SenticNet:** this algorithm exploits the analysis of feelings concept, which means it carries out the polarity detection (Binary classification of the text. It can be positive or negative and varies between 1 and -1) and it realizes the acknowledgement of emotions through semantic Web. The main goal is to give meaning to the contents published in the internet, in order, to be understood by both human and machine [11].
- **PANAS:** Is a method for analysing feelings, which has two mood scales. One measures the positive affection and the other the negative affection. This algorithm does not work online as the previous ones, but it does work using a questionnaire filled by the users [5].
- **SentiStrength:** according to [6] this algorithm has several lists of negative and positive words and also an emoticons list with the respective punctuations. The SentiStrength analyses such a whole text as word by word and the punctuation is from 1 to 5 if it is positive and from -1 to -5 if it is negative. For example, the word “like” have a value of 2 and “love” is value higher, which means, a 3. So, we can conclude that “love” has a stronger sentiment than “like”. It contains a dictionary in almost every language and is available in Windows program and Java code.

To be possible to verify which one of the algorithms would be more suitable to our work, we hold a scale to compare each one according to several parameters (Figure 1), as if it contained or did not contained Portuguese language; if it was or if it was not available free of charge, in order to have a more accessible use; if it evaluated short or long texts; if it placed the words in categories, which means, if it divided them in positive and negative feelings and finally, if it evaluated emoticons.

Through the scale and the research conducted, the SentiStrength algorithm was chosen to carry out the feelings analysis in this project. Professor Mike Thelwall created this algorithm and it is considered to be the best feelings classifier. It has already been used in several important events, such as, The Olympic Games in London, in 2012, and Super Bowl in 2014.

SentiStrength was the one that we choose, as it analyses emoticons which is quite important because, nowadays many people use them to express feelings. It is available free of charge which leads to detect feelings easier since any user can

benefit from its services. It analyses short texts but this may be adapted to analyse long texts as we will see in the next topic. Finally, it also separates the words into categories, from positive to negative feelings. The tests carried out were only done in the SentiStrength because the other algorithms are not available for its use.

Algorithms	Portuguese Language	Free	Tap of text	Categorize	Emoticons
<i>Emoticons</i>		X			X
<i>LJWC</i>	X		Long	X	
<i>SentiWordNet</i>	X		Short	X	
<i>SenticNet</i>			Short	X	
<i>PANAS</i>	X	X	Short	X	
<i>SentiStrength</i>	X	X	Short	X	X

Fig. 1. Comparative table.

3 Person's Feelings / Written Comments

A person's feeling is expressed through a written comment about what they feel concerning a particular subject. The sentiment analysis is the analysis of each comment. The goal is to obtain people's opinion produced by their feelings about a subject. They present those feelings through the words they use when they make comments on written form. Then, the newspaper 'owner' will analyze it and draw some conclusions about the user satisfaction or dissatisfaction about a newspaper narrative. From those results, the user (newspaper) can improve or change the way future narratives will be written or can collect the data to inform others about the acceptance levels that news produced on the reader. As stated before levels are represented in a quantitative order and each work is quoted according to positive and negative values.

4 Application Description

To continue this project and be able to get an input in order to be analyzed by SentiStrength, several components had to be drawn up as a Web Spider and a filter. Both components were performed using the Python scripting language considering that it is easier to implement, and it has a library called BeautifulSoup, which has functions that allow to extract archive data HTML.

4.1 Web Spider

A Web Spider [14] is a computer program that browses by the World Wide Web in a methodical and automated way. For the development applications of the web spider several frameworks can be used and one of them is the Scrapy [10]. It is an open source framework implemented in Python which provides components for the selection and extraction of data from sources as HTML and XML.

This component was developed to withdraw specific informations of the HTML's components of a particular site, in that case, online newspapers sites. The comments are extracted depending on the date and topic entered by users in order to simplify the research and therefore the information to extract. In such a way to remove the comments according to the conditions imposed by the user, the web spider will compare the introduced dates with every news date; if it is within the range it shall be checked in the title of itself, if it finds the topic introduced by the user. In case all these requirements occur, the comments will be removed and stored in a text file.

The web spider removes the comments through the HTML tags presented in each site and, each of them has different tags. The negative point is that it needs to make different web spiders for different newspapers. In some newspapers the comment section is a Facebook plug-in, this make the withdrawal of the comments harder because the plug-in redirects to another HTML page. So, in this project we do not remove the comments from the newspapers that have this section. In the future we will sort this problem.

4.2 Filter

There is a large amount of repeated information and the SentiStrength [6] only evaluates texts written in one line. It was decided to carry out a filter where all the repeated information's, line breaks and enters are eliminated to get an evaluation more concise of the matter.

Keeping the file initial order resulting from the web spider, the filter begins by withdrawing the repeated comments and then removes the characters “\n” and “\r” so the comments are all kept in just one line and be considered by the SentiStrength as one only text.

4.3 SentiStrength

As referred in section 2 the algorithm SentiStrength was used to analyze feelings resulting from comments removed from the online newspapers. The input for the algorithm was the file resulting from the filter, since this is already without repeated comments and with it in just one line to turn the analysis easier. After the file analysis by SentiStrength the output was also a text file with the general classification of the positive and negative feelings and with the individual classification of each word. Through the general average presented in the document it is possible to determine which feelings predominates in the extracted comments.

4.4 Tests

After the comments extracted using the web spider and filter we can initiate the comments analysis by algorithms. The algorithms we found available to perform the tests were the SentiStrength and the LIWC. Though the LIWC full version is paid, we can find it online with a limited characters version (5000). Therefore, we shaved the sample to allow the analysis using the two algorithms.

SentiStrength: Coursing through a set of comments, a meaning according to these is withdrawn through the words, which composed it. It was verified that if each one of these words were in the dictionary they were classified according to their value, negative or positive. Finally, values of each presented word in each comment were summed up, which reflected an average of the positive and negative values of the comment.

Through these values it was possible to say if the comment contained a positive or negative emotion.

LIWC: In the version that is available on LIWC site only English language texts are analyzed. We can also verify that LIWC has several options to classify the texts that we wish to analyze, such as personal writing, social network, scientific writing among others. LIWC analyze the texts as follows:

The module for the examination compares each text word with the dictionary defined by the user. The dictionary identifies the words, which are associated with important psychologically classes. After all, the words in the text are counted. LIWC calculates the total percentage of words that correspond to each dictionary class. For instances, if the LIWC analyses a 2000 words text we can determine that, in these words there are 150 pronouns and 84 words with positive emotions. It converts these numbers to percentages and the outcome will be 7,5% pronouns and 4,2% words containing positive emotions [4].

LIWC carries out a more detailed analysis in what feelings and words classification concerns. In its free version it has a huge limitation: the number of characters. So, we choose the SentiStrength because its analysis is simpler, and it is available without limitation.

4.5 Interface

The user can take advantage of the components mentioned earlier through a developed web application. Some of the application's storyboards are in next figures.

The user chooses the comments that they want to extract from newspapers. Figure 2 shows the choices that the user can make. To have a personalized search, they can only choose one newspaper or choose several.

After choosing the newspapers, the user chooses the time interval from which comments will be taken.

The search can be more detailed if the user searches directly the event they want. In this case, comments are extracted depending on the topic that the user introduced.

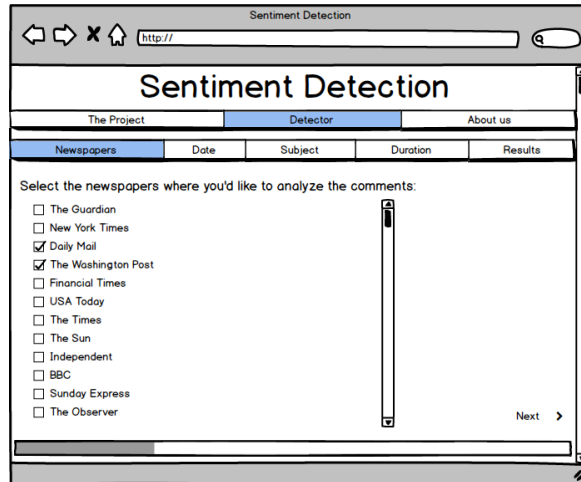


Fig. 2. Select the newspaper.

Since the application takes some time to extract the comments from the newspapers, the user can select the time that he wants to wait or even that, the total time for the application to finish the search.

Figure 3 shows the results obtained through the analysis of comments by the algorithm SentiStrength. The user can see how many comments were analyzed, so they know if the analysis is based on one or many comments. The average of positive and negative feelings in the comments will be displayed as an image, as well as the values that represent those feelings. In the end, it will be possible to verify if the comments expressed more negative or positive feelings.

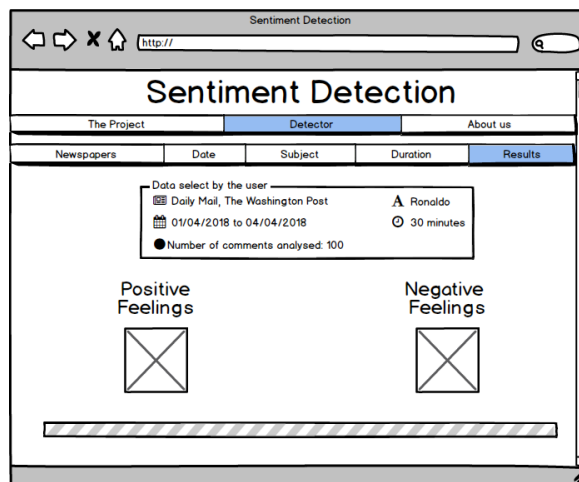
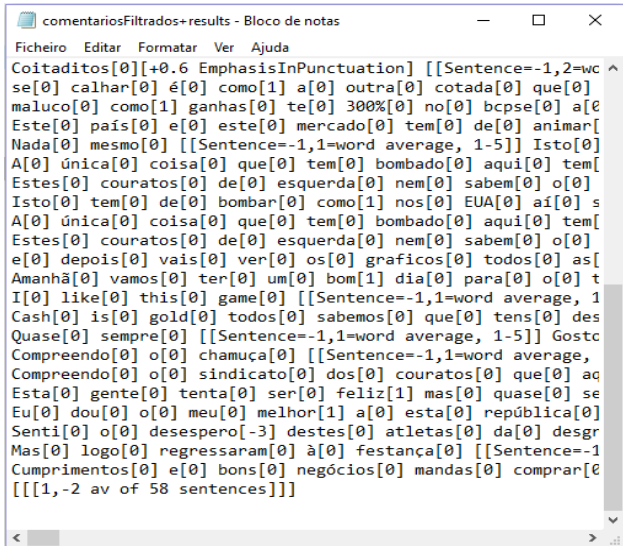


Fig. 2. Select the newspaper.

5 Results

For the results presented below, we searched for news between February 6 and 7, 2018 under the topic "PSI-20". In Figure 4 it is possible to verify the obtained results.



```
comentariosFiltrados+results - Bloco de notas
Ficheiro Editar Formatar Ver Ajuda
Coitaditos[0][+0.6 EmphasisInPunctuation] [[Sentence=-1,2=wc
se[0] calhar[0] é[0] como[1] a[0] outra[0] cotada[0] que[0]
maluco[0] como[1] ganhas[0] te[0] 300%[0] no[0] bcpse[0] a[0]
Este[0] país[0] e[0] este[0] mercado[0] tem[0] de[0] animar[
Nada[0] mesmo[0] [[Sentence=-1,1=word average, 1-5]] Isto[0]
A[0] única[0] coisa[0] que[0] tem[0] bombado[0] aqui[0] tem[
Estes[0] couratos[0] de[0] esquerda[0] nem[0] sabem[0] o[0]
Isto[0] tem[0] de[0] bombar[0] como[1] nos[0] EUA[0] ai[0] s
A[0] única[0] coisa[0] que[0] tem[0] bombado[0] aqui[0] tem[
Estes[0] couratos[0] de[0] esquerda[0] nem[0] sabem[0] o[0]
e[0] depois[0] vais[0] ver[0] os[0] graficos[0] todos[0] as[
Amanhã[0] vamos[0] ter[0] um[0] bom[1] dia[0] para[0] o[0] t
I[0] like[0] this[0] game[0] [[Sentence=-1,1=word average, 1
Cash[0] is[0] gold[0] todos[0] sabemos[0] que[0] tens[0] des
Quase[0] sempre[0] [[Sentence=-1,1=word average, 1-5]] Gostc
Compreendo[0] o[0] chamuca[0] [[Sentence=-1,1=word average,
Compreendo[0] o[0] sindicato[0] dos[0] couratos[0] que[0] aq
Esta[0] gente[0] tenta[0] ser[0] feliz[1] mas[0] quase[0] se
Eu[0] dou[0] o[0] meu[0] melhor[1] a[0] esta[0] república[0]
Senti[0] o[0] desespero[-3] destes[0] atletas[0] da[0] desgr
Mas[0] logo[0] regressaram[0] à[0] festaça[0] [[Sentence=-1
Cumprimentos[0] e[0] bons[0] negócios[0] mandas[0] comprar[0]
[[[1,-2 av of 58 sentences]]]
```

Fig. 4. Results.

To feed the SentiStrength algorithm, comments were taken from a Portuguese newspaper, the *Jornal de Negócios*. Since Portuguese is our native language, we elaborated web spiders for newspapers on this language. The newspaper analysed deals with topics such as economics, finance, and companies, among others.

The PSI-20 is the main reference index of the Portuguese capital market. This subject concerns economy and the opinions are given depending on the state of this index. During the analysis we find that feelings related to this theme are mostly negative. The results are obtainable as an average of the feelings presented in all comments.

6 Conclusions and Future Work

This paper presents the work in progress concerning a systems development for feelings analysis. Authors showed some of the existent algorithms and their characteristics. From those several feelings detectors an algorithm, SentiStrength was considered the most suitable for our project. It analyses short and long texts once adapted to do it; Divides words in categories meaning positive and negative feelings and it also can analyze emoticons. In a manner to feed the algorithm we developed a Web Spider and a filter that allow us to withdraw the comments and filter them withdrawing the repeated ones to ease its analysis.

This project helps users to know what others feel about a particular topic. It may be useful to know what a population's opinion is about, for example, a hospital or hotel and thereby influence other users in making decisions.

Given the fact that the dictionary for our motherly language, Portuguese is incomplete we intend to improve it and as a result contribute to a better functioning. In our future work, we are still developing the application with the already developed components. Web spider and the filter are included in order to permit a more effective application 'use. The prototype is in the user texts' phase. This work can contribute for the inclusion of different people in digital world.

References

1. Araújo, M., Gonçalves, P., Benevenuto, F.: Métodos para Análise de Sentimentos no Twitter. Salvador, Brasil. Retrieved March 15, 2018, from <http://homepages.dcc.ufmg.br/~fabricio/download/webmedia13.pdf> (2013)
2. Cardie, C., Wiebe, J., Wilson, T., Litman, D.: Combining low-level and summary representations of opinions for multi-perspective question answering. In Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, 20–27 (2003)
3. Conglomerates, P.: LIWC. Retrieved March 15, 2018, from LIWC: <http://liwc.wpengine.com/> (2015)
4. Conglomerates, P.: *LIWC*. Retrieved April 04, 2018, from LIWC: <http://liwc.wpengine.com/how-it-works> (2015)
5. Crawford, J. R., Henry, J. D.: The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large. *British Journal of Clinical Psychology*, 43, 245-265 (2004)
6. CyberEmotions. SentiStrength. Retrieved March 15, 2018, from SentiStrength: <http://sentistrength.wlv.ac.uk/> (2013)
7. Dini, L., Mazzini, G. Opinion classification through information extraction. In Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining), 299–310 (2002)
8. Kushal, D., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of WWW, 519–528 (2003)
9. Das, S., Chen, M.: Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA) (2001)
10. Scrapinghub. Scrapy. Retrieved March 15, 2018, from Scrapy: <https://scrapy.org/>
11. SenticNet. Retrieved March 15, 2018, from SenticNet: <http://sentic.net/> (2009)
12. Thakur, D., Gera, T., & Singh, J. (2015). The Senti Strength Calculator: Engineering the Sentiment from the Opinionated Text. In Proceedings 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, 1103-1108 (2015)
13. University, P.: The Trustees of Princeton University. Retrieved March 15, 2018, from WordNet: <https://wordnet.princeton.edu/> (2018)
14. Wolfe, L.: Definition of Spidering and Web Crawlers. Retrieved March 15, 2018, from the balance: <https://www.thebalance.com/what-is-spidering-or-web-crawling-3515704> (2017)
15. Zhang, L., Wang, S., Liu, B.: Deep Learning for Sentiment Analysis: A Survey. eprint arXiv:1801.07883 (2018)