Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Big Data and Automatic Detection of Topics: Social Network Texts

To cite this article: Jesús Silva *et al* 2020 *J. Phys.: Conf. Ser.* **1432** 012073

Big Data and Automatic Detection of Topics: Social Network Texts

Jesús Silva[1], Hugo Hernández Palma[2], William Niebles Núñez[3], Alex Ruiz-Lazaro[4] and Noel Varela[5]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2] Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.
[3]Universidad de Sucre, Sincelejo, Sucre, Colombia.
[4]Universidad Simón Bolívar, Barranquilla, Atlántico, Colombia
[5]Universidad de la Costa, Barranquilla, Atlántico, Colombia

[1]Email: jesussilvaUPC@gmail.com

Abstract. This paper proposes the analysis of the influence of terms that express feelings in the automatic detection of topics in social networks. This proposal uses an ontology-based methodology which incorporates the ability to identify and eliminate those terms that present a sentimental orientation in social network texts, which can negatively influence the detection of topics. To this end, two resources were used to analyze feelings in order to detect these terms. The proposed system was evaluated with real data sets from the Twitter and Facebook social networks in English and Spanish respectively, demonstrating in both cases the influence of sentimentally oriented terms in the detection of topics in social network texts.

1. Introduction

Today, the social networks is recognized and, as a consequence, the increase in the number of users interacting in these networks, which causes the accumulation of large volumes of unstructured textual data. For this reason, social networks are a very important source of information, so it is to be expected that enterprise, researchers, etc., spend time and resources in their study. However, the great accumulation and lack of structure of the texts makes it practically impossible to process and analyze them automatically on a massive scale. This is the reason for previously organizing texts taking into account the subject matter [1].

The detection of topics from unstructured texts allows to organize these texts by subject, which facilitates its later analysis integrated with conventional data. In [2] a multilingual methodology for the automatic detection of topics in textual data is proposed. Through experimentation, the viability of the proposal was demonstrated, although it should be stressed that the results are not good enough when the texts come from social networks. This is because the detection of topics in more elaborate texts (digital libraries, news websites, etc.) is different when texts belong to social networks. In such systems, users express ideas, facts and feelings on any subject using colloquial language, so it is expected that terms appear in the texts with high frequency for expressing feelings related to certain products, services, etc. [3].

Motivated by the previous problem, this paper analyzes the influence of the terms that express feelings in the automatic detection of topics in social networks. For this purpose, a new approach is proposed to improve the methodology for the automatic detection of the main topics present in textual data proposed in [4], which uses data mining techniques, resources related to feeling analysis, and a multilingual knowledge base. The new proposal makes it possible to identify and eliminate sentimentally oriented terms, with the aim of improving the results of the system on social network texts.

The basic idea is to carry out a filtering that eliminates the words that express feelings during the semantic pre-processing stage of the method. To do this, the lexical resources SentiWordNet [5] and WordNet Affect [6] are used separately and together, to compare the results obtained. Four sets of data were used for the experimentation, which belong to the Twitter and Facebook social networks, in English and Spanish.

2. Previews Studies

The detection of topics from large volumes of texts is a widely analyzed topic in the literature from various points of view. These include the use of methods such as classification algorithms, Latent Dirichlet Allocation (LDA) and grouping algorithms, among others. In the case of classification algorithms, it is necessary to have a set of training data to train the classifier, while LDA and grouping algorithms do not require a previously classified corpus.

There are many studies related to the detection of topics through the use of supervised and semi-supervised hierarchical grouping algorithms, not so for unsupervised algorithms. Such are the cases proposed in [7] [8], where the authors propose approaches based on the use of expert information, in order to improve the results in the detection of main topics.

From the unsupervised point of view, [9] presents a proposal for the automatic detection of topics in textual data based on ontologies. The lexical resource named WordNet Domains [10] was used in order to homogenize the syntactic representation of the concepts in the texts and thus considerably reduce the dimensionality of the problem. The Reuters-21578 data set, which contains texts related to news publications, was used for the experimentation. The results show the viability of the proposal, where the Silhouette Coefficient values are better when the proposed methodology is applied. It should be noted that, although Reuters-21578 texts are real data, they are long, well elaborated and also belong to a restricted domain. If one considers that texts in social networks are short texts and users mainly express their feelings on a given topic, these texts should be pre-processed in a different way in order to extract the present topics.

In social networks, topic detection has been widely used for textual data analysis. Many solutions have appeared for textual analysis in social networks, such as feeling analysis [11], content filtering [12], [13], user interest modeling [14], as well as interest event tracking [15], [16]. In [17], a comparison is made between the content of Twitter texts and a traditional means of communication, the New York Times. For this purpose, unattended topic modeling using the Twitter-LDA model is used to discover these topics in short messages.

On the other hand, several works present a model where the detection and analysis of topics merge with the analysis of feelings [18]. In all of them, the detection of topics is carried out without considering the influence that the terms have with a certain sentimental orientation in that task.

This paper analyses the influence of sentimentally oriented terms in the detection of topics in social networks. For this purpose, a filter is applied during semantic pre-processing with the aim of eliminating the terms that express feelings, which can introduce noise in the detection of topics. This proposal is totally new since, unlike the mentioned works where the detection of topics is merged with the analysis of feelings, in this case what is done is to discard the terms related to feelings, remaining only the terms that provide useful information for the automatic detection of the main topics.

3. Methods

As mentioned above, this paper analyzes the influence of sentimentally oriented terms on the automatic detection of topics on social networks. Bearing in mind that textual data from social networks are written in a more colloquial way, and users tend to express their feelings and opinions about certain products, services, entities, attributes, etc., it is useful to detect and eliminate those terms with a certain sentimental orientation, since such terms do not provide useful information for the detection of topics.

A summary of each of the phases of the proposed system is presented below, for more details see [19]. For the specific case of the semantic pre-processing phase, the filter that allows to identify and discard the terms related to feelings is highlighted.

3.1 Syntactic pre-processing

First, the processes of grammatical category labeling and entity recognition are executed with the Stanford POS [19] and Stanford NER tools, respectively [20]. Then, the tokenization filter is applied and the necessary filters are applied to eliminate the terms that belong to the set of empty words, those that are not identified as nouns by the grammatical tagger, those that are identified as nouns by the entity identifier, as well as those terms that are not found in the Multilingual Central Repository (MCR) knowledge base [21], since all of them do not provide useful information for the detection of topics

3.2 Semantic pre-processing

Once the texts are syntactically pre-processed, the semantic analysis is carried out. In this case, the objective of semantic pre-processing is to homogenize the syntactic representation of the concepts present in the text. The aim is to replace the WordNet Domains [7] tags with which the WordNet senses are tagged by the terms present in the original texts. As already mentioned, the study analyzes the influence of sentimentally oriented terms in the detection of topics. For this reason, Section IV deeply explains the process related to the identification and elimination of sentimentally oriented terms.

3.3 Hierarchical Grouping

Once the texts are homogenized, the hierarchical grouping of the texts is carried out from the WordNet Domain labels. The approach proposed in [1] is used to represent the characteristics. In this paper, only the hierarchical grouping algorithm Complete Link is analyzed using the distance of the cosine as a similarity measure.

3.4 Group labelling

When the grouping phase ends, the process of selecting labels from the groups is carried out, which is a very important task, especially in applications related to data analysis, where the end user needs to know what a particular group is about [20]. In the present study, the Arithmetic Mean is used to determine the most relevant labels of each group of texts.

4. Experimentation

Following, the experimentation will allow to demonstrate the validity of the proposal.

4.1 Data sets

Four data sets, Table 1, belonging to Twitter and Facebook Social Networks were selected. The Twitter data were obtained from the training set in CSV format and consists of seven fields, including the text of the tweets which will be used in this paper. It should be mentioned that these data were selected because they are oriented to the Analysis of Feelings, and constitute a source of great importance for the experimentation.

On the other hand, the Facebook database offers a total of 72 tables. The information collected is all information related to the user's personal and affiliation data, as well as, the interactions they make in their profile and with other users. In this case, Comments of the users are included.

It should be mentioned that besides the language, the texts from Twitter differ with those from Facebook in the following aspects:

Facebook texts, although they deal with different topics, mainly constitute subjects related to the university environment, since the users selected from this social network are students of a university.

The length of the Twitter texts is restricted to 140 characters, while Facebook texts have no restrictions.

Table 1. Description of the Experimental Data Sets

Set	Number of documents	Source	Language	Intervention	# different terms	Total terms
1	80000	Twitter	English	Tweet	4568	20.369
2	200000	Twitter	English	Tweet	6985	254.325
3	80000	Facebook	Spanish	Comment	3541	14.352
4	200000	Facebook	Spanish	Comment	4587	377.587

4.2 Evaluation

In this section, the procedure for evaluating the functioning of the method for the automatic detection of topics is explained by applying the filter to eliminate the terms that express feelings with the different resources and without applying it. For this purpose, the Silhouette Coefficient [22] 1 was used as a measure. This measure makes it possible to analyze the quality of the groups created by the hierarchical grouping algorithms.

The experimentation was carried out using the Complete Link method and making cuts for the following quantities of groups (18, 30, 44, 64, 80, 110 and 160), and for each case, the value of the Silhouette Coefficient was determined when the filter is not applied to eliminate feelings, as well as when it is applied making use of the different resources.

5. Results & Discussions

Tables 2, 3 and 5 show the Silhouette Coefficient values of each experimental set and the different quantities of groups, both when the filter is not applied to eliminate the terms that express feelings (NF) and when it is applied using SentiWordNet (SWN), WordNet Affect (WA) and the combination of both (SWN-WA). It is possible to note that, better results are obtained when the filtering of feelings is applied than when it is not (values in bold).

Table 2. Silhouette Coefficient of the Set 1

Resource/ Groups	18	30	44	64	80	110	160
Express feelings	0.01	0.08	0.20	0.23	0.23	0.22	0.19
SentiWordNet	0.02	0.2	0.17	0.19	0.25	0.26	0.23
WordNet Affect	0.01	0.078	0.16	0.17	0.21	0.24	0.21
combination SWN-WA	0.015	0.068	0.23	0.25	0.23	0.24	0.23

Table 3. Silhouette Coefficient of Set 2

Resource/ Groups	18	30	44	64	80	110	160
Express feelings	0.03	0.068	0.11	0.19	0.32	0.18	0.19
SentiWordNet	0.07	0.078	0.12	0.21	0.36	0.32	0.36
WordNet Affect	0.06	0.025	0.14	0.22	0.28	0.32	0.24
combination SWN-WA	0.05	0.085	0.14	0.23	0.24	0.17	0.23

Table 4. Silhouette Coefficient of Set 3

Resource/ Groups	18	30	44	64	80	110	160
Express feelings	0.14	0.23	0.33	0.44	0.39	0.39	0.40
SentiWordNet	0.18	0.24	0.41	0.49	0.47	0.53	0.50
WordNet Affect	0.24	0.22	0.36	0.43	0.45	0.41	0.41
combination SWN-WA	0.45	0.26	0.41	0.49	0.55	0.54	0.50

Table 5. Silhouette Coefficient of Set 4

Resource/ Groups	18	30	44	64	80	110	160
Express feelings	0.18	0.21	0.33	0.40	0.45	0.44	0.41
SentiWordNet	0.14	0.24	0.47	0.52	0.55	0.55	0.51
WordNet Affect	0.15	0.25	0.40	0.44	0.39	0.44	0.43
combination SWN-WA	0.17	0.25	0.40	0.52	0.51	0.50	0.50

Figures 1 and 2 show the graphs of the Silhouette Coefficient with respect to the number of groups and the resource used to detect the terms with sentimental orientation respectively, using in both cases the Kruskal-Wallis test [23]. The conclusions of the analysis are summarized below:

From 72 groups onwards the Silhouette Coefficient values are stabilized, showing significant differences with the previous quantities.

There is a remarked difference between the results obtained with the SWN and SWN-WA resources, and those obtained with the WA resource and when no filter is applied to detect terms with sentimental orientation; although there are no significant differences between SWN and SWN-WA if considering that the use of SWN-WA entails the use of another lexical resource.

On the other hand, Figure 3 shows the graph of the Silhouette Coefficient with respect to the social network from which the texts come. The Wilcoxon test [24] was carried out and it can be concluded that there is a big difference for the social networks used, as Facebook provides the best results. This is largely due to the fact that a large part of the users of this network belong to a university context, so that the conversation domain is more restricted.

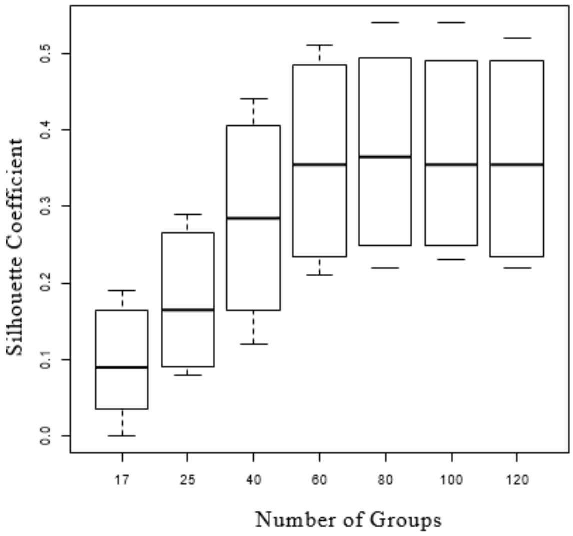

Figure 1. Graph between groups and the Silhouette Coefficient

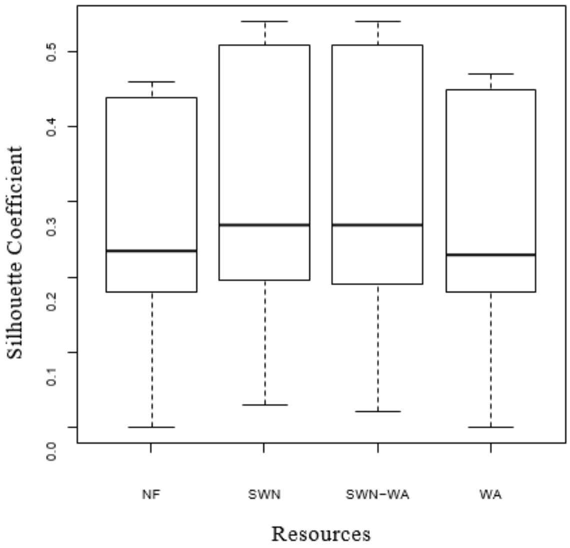

Figure 2. Graph between the Silhouette Coefficient and the resources used

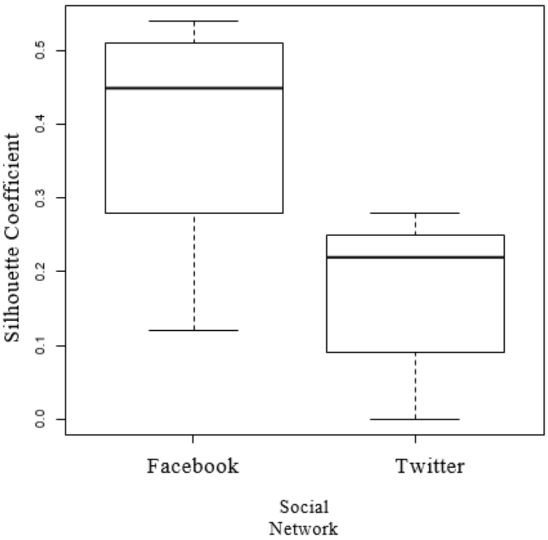

Figure 3. Graph between the Silhouette Coefficient and the social networks used

6. Conclusions

The experiments carried out with both Twitter and Facebook to show the viability of the system. Experiments were performed without applying the filter to eliminate feelings and then applying it with two lexical resources related to the analysis of feelings. In each case, cuts were made in the quantities of groups already mentioned and the Silhouette Coefficient was calculated. The results achieved when the filter was applied improve the results obtained when the filter is not applied. The resource with which a better performance was obtained was SentiWordNet, since, although the combination of both resources improves SentiWordNet in certain cases, the differences are not significant considering that all the concepts of WordNet Affect are incorporated.

References

[1] A. Gonzalez-Agirre, E. Laparra, y G. Laparra, “Multilingual central repository version 3.0,” in Proceedings of the Eight International Con- ference on Language Resources and Evaluation (LREC’12). Istanbul, Turkey: European Language Resources Association (ELRA), may 2012

[2] P. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” J. Comput. Appl. Math., vol. 20, no. 1, pp. 53–65, Nov. 1987. [Online]. Disponible: http://dx.doi.org/10.1016/0377- 0427(87)90125-7.

[3] F. Wilcoxon, “Individual comparisons by ranking methods,” Biometrics Bulletin, vol. 1, no. 6, pp. 80–83, 1945.

[4] K. Toutanova, D. Klein, C. D. Manning, y Y. Singer, “Feature-rich part- of-speech tagging with a cyclic dependency network,” in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, ser. NAACL ’03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 173–180. [Online]. Disponible: http://dx.doi.org/10.3115/1073445.1073478.

[5] Lis-Gutiérrez JP., Gaitán-Angulo M., Henao L.C., Viloria A., Aguilera-Hernández D., Portillo-Medina R. (2018) Measures of Concentration and Stability: Two Pedagogical Tools for Industrial Organization Courses. In: Tan Y., Shi Y., Tang Q. (eds) Advances in Swarm Intelligence. ICSI 2018. Lecture Notes in Computer Science, vol 10942. Springer, Cham

[6] W. X. Zhao, J. Weng, J. He, E.-P. Lim, y H. Yan, “Comparing twitter and traditional media using topic models,” in 33rd European conference on advances in information retrieval (ECIR11). Berlin, Heidelberg: Springer-Verlag., 2011, pp. 338–349.

[7] Viloria, A., & Gaitan-Angulo, M. (2016). Statistical Adjustment Module Advanced Optimizer Planner and SAP Generated the Case of a Food Production Company. Indian Journal Of Science And Technology, 9(47). doi:10.17485/ijst/2016/v9i47/107371.

[8] F. Villada, N. Muñoz, y E. García, Aplicación de las Redes Neuronales al Pronóstico de Precios en Mercado de Valores, Información tecnológica, vol. 23, núm. 4, pp. 11–20. 2012..

[9] N. Sapankevych y R. Sankar, “Time Series Prediction Using Support Vector Machines: A Survey”, IEEE Computational Intelligence Magazine, vol. 4, núm. 2, pp. 24–38, may 2009.

[10] N. Swanson y H. White, “Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models”, International Journal of Forecasting, vol. 13, núm. 4, pp. 439–461, 1997.

[11] E. M. Toro, D. A. Mejia, y H. Salazar, “Pronóstico de ventas usando redes neuronales”, Scientia et technica, vol. 10, núm. 26, 2004.

[12] Hernández, J. A., Burlak, G., Muñoz Arteaga, J., y Ochoa, A. (2006). Propuesta para la evaluación de objetos de aprendizaje desde una perspectiva integral usando minería de datos. En A. Hernández y J. Zechinelli (Eds.), Avances en la ciencia de la computación (pp. 382-387). México: Universidad Autónoma de México.

[13] Romero, C., Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications, 33(1), 135-146.

[14] Romero, C., y Ventura, S. (2010). Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40(6), 601-618. Disponible en: http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?reload=true&punumber=5326

[15] Choudhury, A. and Jones, J. Crop yield prediction using time series models, Journal of Economics and Economic Education Research., 15, 53-68, 2014.

[16] Scheffer, T. (2004). Finding Association Rules that Trade Support Optimally Against Confidence. Intelligent Data Analysis, 9(4), 381-395.

[16] Ruß G. Data Mining of Agricultural Yield Data: A Comparison of Regression Models, In: Perner P. (eds) Advances in Data Mining. Applications and Theoretical Aspects, ICDM 2009. Lecture Notes in Computer Science, vol 5633.

[17] Viloria A., Lis-Gutiérrez JP., Gaitán-Angulo M., Godoy A.R.M., Moreno G.C., Kamatkar S.J. (2018) Methodology for the Design of a Student Pattern Recognition Tool to Facilitate the

Teaching - Learning Process Through Knowledge Data Discovery (Big Data). In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham

[18] Y. Rao, Q. Li, X. Mao, y L. Wenyin, "Sentiment topic models for social emotion mining," Information Sciences, vol. 266, pp. 90 – 100, 2014. [Online]. Disponible: http://www.sciencedirect.com/science/article/pii/ S002002551400019X

[19] K. Gutiérrez-Batista, J. R. Campaña, M.-A. Vila, y M. J. Martin- Bautista, "An ontology-based framework for automatic topic detection in multilingual environments," International Journal of Intelligent Systems, vol. 33, no. 7, pp. 1459–1475, 2018. [Online]. Disponible: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.21986

[20] J. Wu, W. Gao, B. Zhang, J. Liu, y C. Li, "Cluster based detection and analysis of internet topics," in 4th International Symposium on Computational Intelligence and Design, ISCID 2011, vol. 2, 2011, pp. 371–374.

[21] L. Zheng y T. Li, "Semi-supervised hierarchical clustering," in Proceedings of the 2011 IEEE 11th International Conference on Data Mining, ser. ICDM '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 982–991. [Online]. Disponible: http://dx.doi.org/10.1109/ICDM.2011.130

[22] C. Lin y Y. He, "Joint sentiment/topic model for sentiment analysis," in 18th ACM Conference on Information and Knowledge Management 8CIKM09). New York, NY, USA: ACM, 2009, pp. 375–384.

[23] J. Duan y J. Zeng, "Web objectionable text content detection using topic modeling technique," Expert Systems with Applications, vol. 40, pp. 6094–6104., 2013.

[24] M. Pennacchiotti y S. Gurumurthy, "Investigating topic models for social media user recommendation," in 20th International Conference Companion on World Wide Web. New York, NY, USA: ACM, 2011, pp. 101–102.