Machine Learning Algorithms and Techniques for Sentiment Analysis in Scientific Paper Reviews: A Systematic Literature Review

Samuel Machado, University of Minho, Portugal, samuel.c.machado98@gmail.com
Ana Carolina Ribeiro, University of Minho, Centro Algoritmi, Portugal, anacfr1@hotmail.com
Jorge Oliveira e Sá, University of Minho, Centro Algoritmi, Portugal, jos@dsi.uminho.pt

Abstract

Sentiment analysis also referred to as opinion mining, is an automated process for identifying and classifying subjective information such as sentiments from a piece of text usually comments and reviews. Supported by machine learning algorithms, it is possible to identify positive, neutral or negative opinions, being possible to rank or classify them in order to reach some kind of conclusion or obtain any type of information. Thus, this paper aims to perform a systematic literature review in order to report the state-of-the-art of machine learning techniques for sentiment analysis applied to texts of reviews, comments and evaluations of scientific papers.

Keywords: Machine Learning; Sentiment Analysis; Opinion Mining; Systematic Literature Review

1. INTRODUCTION

About 6,849.32 journal articles are published every day, distributed across all areas of research, and this number has been constantly increasing (Ware & Mabe, 2015). Therefore, research studies and projects have become more complex due to the huge amounts of papers and information which results in spending more time than required trying to find the scientific paper or papers that will provide the information needed to accomplish the research. Besides that, the lack of criteria to help researchers finding what they are looking for turns researches into non-friendly processes. Therefore, and without any other references to filter the papers besides the number of citations (the only official and most common criteria), other problems like the lack of quality and relevance in the obtained information appear and not only do they not help in the research process, but they become counterproductive too.

Thus, is visible the need for auxiliary solutions that complement existing methods such as the imposition of altmetrics, the most widely used term to describe alternative assessment metrics. Those metrics are an alternative to the established citation counts and usage stats—and/or metrics about alternative research outputs (Erdt, Nagarajan, Sin, & Theng, 2016).

One way to achieve those classifications systems can be through the social experience of researchers that they share under the form of unstructured text obtained from comments, reviews, shares or

ISSN 2183-489X

classifications to those papers. By developing a sentiment analysis classification system that analyzes that information with the purpose of assigning them a label which indicates that the comment is positive or negative, it is possible to classify the paper itself with those same labels.

Thus, this paper aims to provide an overview of the state-of-the-art in machine learning algorithms and techniques for sentiment analysis through the review of the most recent studies and reviews that have been conducted in Sentiment Analysis as well as identifying the most suitable algorithms to perform sentiment analysis classification on the scientific paper comments.

The structure of this paper is composed as follows: Research Method, Sentiment Analysis, Literature Review Results and Conclusions.

2. RESEARCH METHOD

The method used to structure the research process and the write of this literature review were based in the guidelines provided by the paper "Analyzing the Past To Prepare for the Future: Writing a Literature Review" (Webster & Watson, 2002).

2.1. Initial Literature Search

This research used two search engines: Scopus and ScienceDirect. The process started with the identification and definition of the initial search criteria and terms.

Due to the high number of results obtained with the first defined search criteria and search terms and due to the project limitations, was necessary to refine the search terms and the search criteria in order to obtain a lower number of results. Thus, and after several filtering processes were defined the final criteria for this stage, i.e, *Year >= 2018* and *Type = Review* were the final search terms and some of the initial search terms were set aside due to their extensiveness. Table 1 synthesizes the obtained results of the search stage.

SEARCH TERM	SCIENCE DIRECT	SCOPUS
1. machine learning in sentiment analysis	65	7
2. machine learning in sentiment and emotional analysis	22	0
3. machine learning algorithms and techniques for sentimental and emotional analysis	2	0
4. opinion mining for sentiment analysis	26	14
5. opinion mining for sentiment and emotions analysis	12	0
6. sentiment analysis and opinion mining applied to reviews	25	0
7. sentiment and emotion analysis and opinion mining applied to scientific reviews	6	2
8. "sentiment analysis" and "opinion mining"	11	12

Table 1 – Total number of reviews

2.2. Initial Literature Selection

The initial selection is based on a preliminary reading of title, abstract, keywords, and introduction of all papers, with the goal of identifying the paper relevance. Table 2 presents the comparison between the selected reviews and the obtained due to the existence of repeated papers in multiple terms.

SEARCH TERM	SCIENCEDIRECT			SCOPUS		
	Selected	**Repeated**	**Obtained**	**Selected**	**Repeated**	**Obtained**
1.	-	-	-	7	5	3
2.	-	-	-	-	-	-
3.	2	2	0	-	-	-
4.	26	4	4	14	8	3
5.	12	3	2	-	-	-
6.	25	4	4	-	-	-
7.	6	3	0	2	1	0
8.	11	4	1	12	7	3
Total	82	20	**11**	35	21	**9**

Table 2 – Number of obtained and repeated reviews

2.3. Backward Tracking Process

The purpose of this backward tracking process is to complement the research as well as fix any informational gaps that might exist in the current stage of the research by obtaining additional relevant papers that the selected ones have made reference to. This can be accomplished by building a cross-reference matrix between reviews in order to determine the frequency of each reference present in the obtained reviews. However, due to the huge number of references that the selected papers have (over 1500) and how much inconsistent they are (different authors have different types of making references), it would be tough to make this cross-reference table manually.

Thus, was developed a simple application that returns the number of occurrences of each reference from the complete given references list. To develop this application was made a preliminary analysis of the references list in order to detect which adjustments were needed in the data. This process resulted in 13 more papers to complement the ones obtained previously.

3. SENTIMENT ANALYSIS

According to the data of Scopus, sentiment analysis is a field that has been under research since 1997. In the last 14 years (2005-2019), the number of published papers related to the sentiment analysis (10643 papers according to Scopus) increased a lot compared with the papers published before 2005 (8 papers according to Scopus). If we take a look only at the last 14 years, we can see that it still represents 99.93% of the total number of papers that have been published in this field

until now. Figure 1 and Figure 2 represent the fast increase in the number of papers in sentiment analysis.

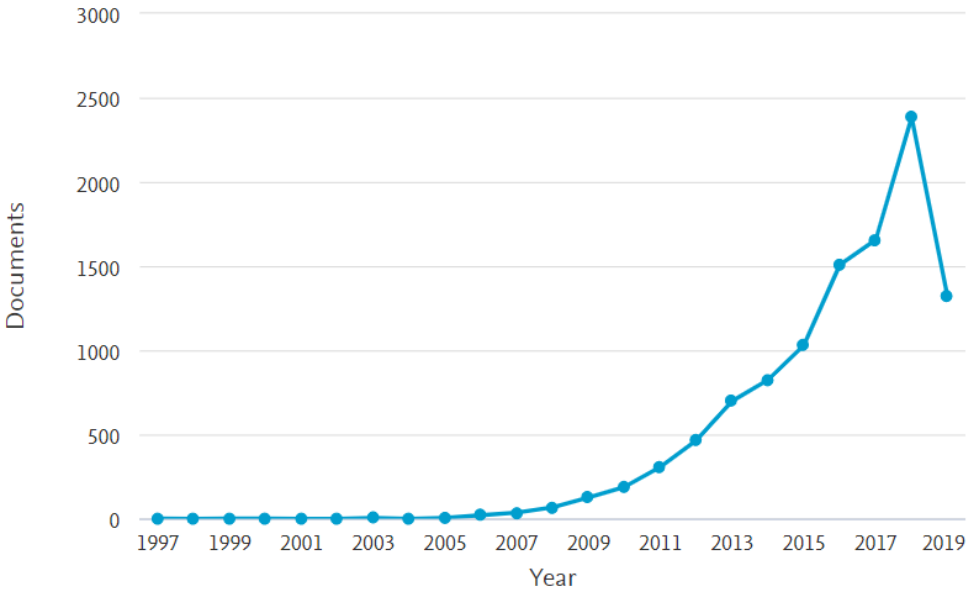

Figure 1 - Number of "sentiment analysis" documents published per year. Source:Scopus.

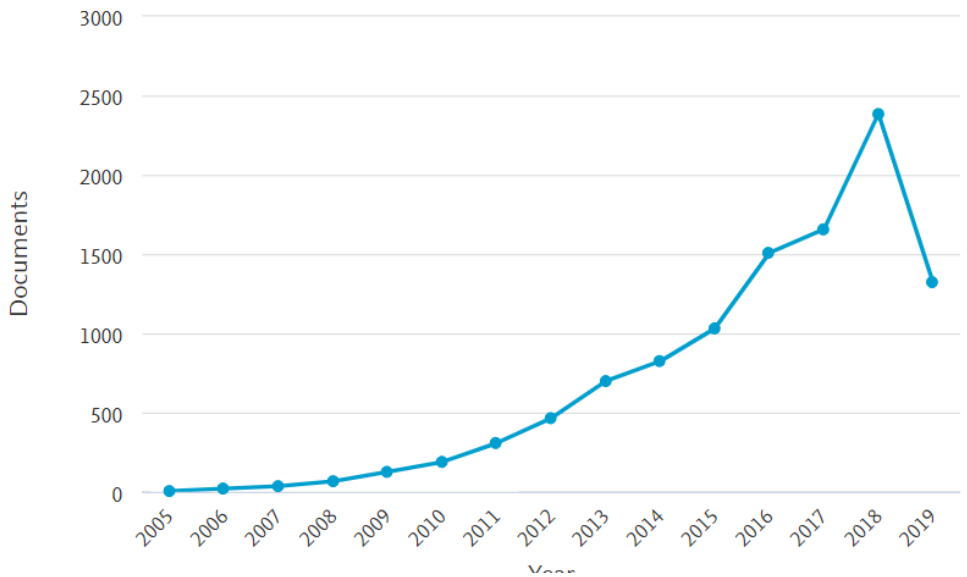

Figure 2 - Number of "sentiment analysis" documents publised in the last 8 years. Source:Scopus.

Although the results from Scopus only provide papers after 1997, there are studies that make reference to papers related to sentiment analysis from 1940 (Mäntylä, Graziotin, & Kuutila, 2018).

So why the sudden interest in this area even if it exists for such a long time? This field of research can have various applications such as detection of hate speech (Fortuna & Nunes, 2018), natural environment monitorization (Lei, Marfia, Pau, & Tse, 2018) or even detecting terrorism-related publications (Choi et al., 2014). However, the main reason could be the emergent need that the organizations are facing to answer the impositions and suggestions made by customers in order to retain their market share and, if possible, increase it.

In the context of this research, predicting the paper relevance based on the reviews made by the existing readers could be a good way to help clean and organize the search process of the researchers. We can think of this as the same as a customer buying a product. If he has any source of information that the product isn't good, he will not be buying it. Unfortunately, sometimes a published paper doesn´t have the minimum quality requirements and probably could mislead researches to some topics that are completely off from the starting point.

Thus, and before a deep exposure of the components that constitute the sentiment analysis, it is important to have a clear understanding of what really is sentiment analysis.

Sentiment analysis, often related to opinion mining, opinion summarization, and text mining can be defined as the use of machine learning to identify and extract subjective information in a piece of writing (Boudad et al., 2018). It also can be defined as a data mining technique that systematically evaluates attitudes and opinions on a topic of interest using machine learning techniques (Rambocas & Gama, 2013) or a process of automatically summarizing opinions that are related to the same topic (Moussa, Mohamed, & Haggag, 2018) in order to retrieve any type of information that could contribute actively in the decision making process. Such information provided by the analysis of the user's experiences, in case of a product, can influence future customers decisions resulting in purchases or reverting the decision of buying the product (Heydari et al., 2015). From the organization point of view, positive or negative reviews can have different impacts on the sales process such as financial gains and/or an increase in the number of customers if positive or sales loss if negative (Tavakoli, Zhao, Heydari, & Nenadić, 2018).

In the case of scientific papers, machine learning algorithms can be used in order to classify them. However, the performance of the machine learning algorithms can be improved by selecting the most suitable algorithms for the given context. Thus, in order to improve the results obtained it is important to frame the context into three different features of sentiment analysis.

3.1. Sentiment Analysis Levels

It is possible to identify three different levels in sentiment analysis. These levels are related to the granularity of the data (Boudad et al., 2018) and the can provide different levels of complexity due to different levels of data preparation. The three different levels are document level, sentence level, and aspect level.

Document level granularity tries to predict the overall sentiment polarity of a whole paragraph or document (Mäntylä et al., 2018) supposing that the whole document expresses an opinion to only one entity with only one opinion holder (Boudad et al., 2018). This sentiment analysis level is not viable to apply in case the document has opinions in more than one entity.

The sentence level granularity tries to identify the sentiment polarity expressed in each sentence (Sundermann, Domingues, Sinoara, Marcacini, & Rezende, 2019). This level of analysis reveals itself to be more challenging than the previous one because the sentiment polarity expressed in each sentence may depend on the context in which it is inserted.

Aspect level granularity it is considered to be a finer-grained analysis (Boudad et al., 2018) since it tries to identify different aspects of the entity under review and also predict the polarity related to every aspect identified. For example, a review on a cellphone (entity) can mention different components of the cellphone such as screen, battery, etc (aspects). In this type of analysis, the entity can have a negative classification having aspects classified positively.

3.2. Sentiment Analysis Approaches

Similarly to the sentiment analysis levels, there are also three general sentiment analysis approaches: supervised approach, unsupervised approach, and hybrid (semi-supervised) approach.

The supervised approach requires a set of labeled data to perform the algorithm training process. This data, besides the input values, must have also the output values to improve the results obtained through the algorithms parameters calibration (Zimbra, Abbasi, Zeng, & Chen, 2018). The evaluation of these algorithms results from a comparison between the obtained results and the label assigned.

On the other hand, the unsupervised approach tries to predict the sentiment polarity of a review by making use of the words and the associated sentiment (Boudad et al., 2018). It can resort to words lexicons that contain combinations between words and the respective sentiment in order to determine the overall existent sentiment.

Another more recent approach is the hybrid approach also called semi-supervised approach or weakly supervised approach (Boudad et al., 2018). This approach results from the combination of the two other approaches in order to try to obtain better results. This approach uses a set of labeled data to train the algorithms and unlabeled data to test the algorithm in order to develop classification systems the closest possible to the reality.

3.3. Sentiment Analysis Classification

In terms of sentiment analysis classification, there are also three commonly classification types adopted: classification, multi-class classification, and multi-label classification.

Binary classification provides a type of sentiment polarity classification based in binary labeling that assigns the values 1 and 0 usually to positive and negative respectively.

The multi-class classification labels the data according to a scale of classes based on the strength of the expressed sentiment. Instead of having only positive and negative as classes we can have a list

of defined of classes such as extremely negative, negative, neutral, positive, and extremely positive (Boudad et al., 2018).

Multi-label classification is slightly different from the other two since it classifies data by assigning them one or more labels. This can be useful for identifying not the polarity of the sentiment but the sentiment itself. With this classification system its possible to identify different sentiments present in the data assigning a label for each different identified sentiment.

3.4. Sentiment Analysis Selection

After analyzing the different features of sentiment analysis, were selected the ones that were going to be adopted in the classification system development. This selection is important in order to restrict the algorithms search to those who fit the project needs.

Therefore, in the sentiment analysis levels, both document level and sentence level revealed themselves viable to adopt in the classification of reviews from scientific papers since the obtained data fulfills the defined requirements. Even though a review to a scientific paper can be made to an aspect level, for example, "The keywords are not aligned with the paper subject." (having keywords as an aspect), the goal of this project is to not classify every aspect of the review but just obtain a classification to the whole review. Furthermore, those kinds of reviews are much easier to find in product reviews than in scientific papers reviews.

For the sentiment analysis approaches, unsupervised approach and the hybrid approach seems to be the most viable approaches in the classification of reviews from scientific papers since the set of obtained data is unlabeled. Finding a good set of labeled data proved to be a really tough task and the manual labeling of the complete data seems to be unviable due to resources limitations. Therefore, partially label the data to apply a hybrid approach or the adaptation of an existent lexicon are the best options.

In the case of the sentiment analysis classification, binary classification seems to be the one which is more aligned with this project initial goals (predict if the review is either positive or negative - 2 classes).

Figure 3 synthesizes the selections made on each feature.

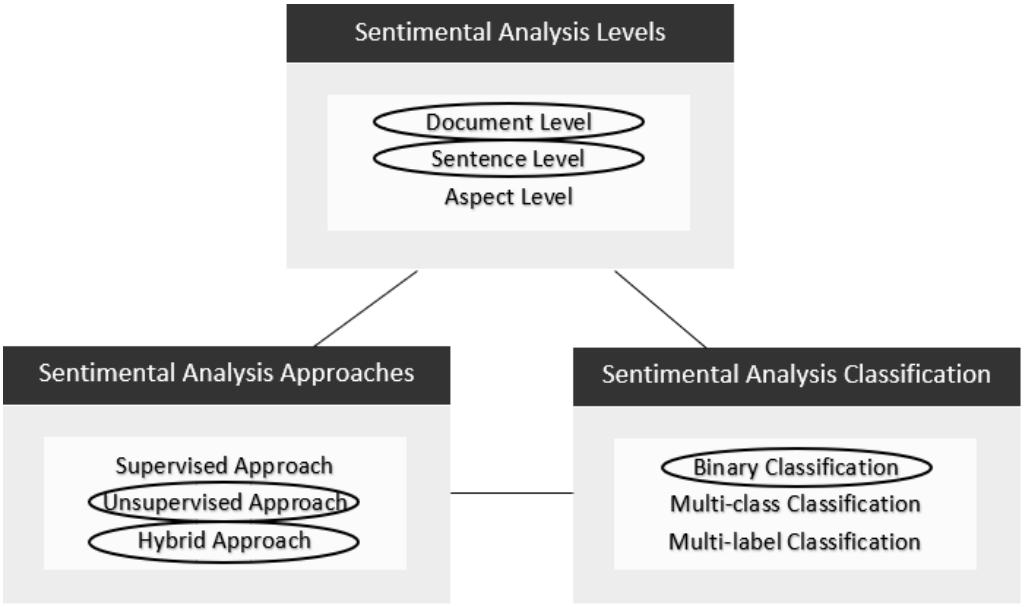

Figure 3 - Sentiment analysis selected features.

3.5. Sentiment Analysis Challenges

Sentiment analysis still faces a lot of challenges associated with the different stages of the process:

- The fast evolution of the language especially between young communities (Fortuna & Nunes, 2018);
- The existence of synonyms, in aspect based levels, can cause different aspects to be analyzed separately when they should be treated as the same aspect (Afzaal, Usman, Fong, & Fong, 2019);
- Word sense ambiguity and language-specific structures (Do, Prasad, Maag, & Alsadoon, 2019);
- Retrieve relevant data (Moussa et al., 2018);
- Data interpretation; and
- Unstructured data (Tavakoli et al., 2018).

The data pre-processing step can itself be very challenging due to the huge inconsistency of the data as well as subjective forms of expression such as sarcasm or irony. The listed challenges will be approached in future work.

4. LITERATURE REVIEW RESULTS

The present section aims to summarize the results obtained from the review of the obtained literature in order to determine the best machine algorithms to the sentiment analysis classification.

As previously said, the presented results will be oriented to the selected components of each feature in order to determine the algorithms that produce best results when faced with problems with similar characteristic than the problem under study.

Figure 4 proposes the algorithms and models for the sentiment classification of scientific paper reviews that revealed the best values of accuracy in the analysis of every obtained review.

		Sentimental Analysis Approaches		
		Supervised	**Unsupervised**	**Hybrid**
Sentimental Analysis Levels	**Document**	SVM NB K-NN	Based-Lexicon Latent Dirichlet Allocation	SVM NB
	Sentence	SVM NB	Based-Lexicon	Linear Regression

Figure 4 - Selected machine learning algorithms.

For the supervised approach is very common the implementation of SVM and NB since they are two most well-known algorithms not only for sentiment analysis but for prediction tasks in general. Even though they exist since the nineteen-sixties, they still produce really satisfactory results. For hybrid approaches, the SVM and NB algorithms still provide very good results such as 98.31% as the best result for a binary classification system (Al-amrani, Lazaar, Eddine, & Kadiri, 2018).

In the other hand, based-lexicon models are the most explored solution for unsupervised approaches. Examples of results obtained by implementing a solution lexicon based are 74.65% of accuracy (Musto, Semeraro, & Polignano, 2014) or 93.23% as the higher value of precision (Fernández-Gavilanes, Àlvarez-López, Juncal-Martínez, Costa-Montenegro, & González-Castaño, 2015) in different contexts.

However, the results to be obtained from the implementation of the selected methods could not perform as well as they perform in those studies. The accuracy of the algorithms may vary depending on the available data that is being passed as input, the type of pre-processing which the data have been exposed to or even the algorithms configuration itself. These are only some examples of factors that can affect the performance of the classification system.

Thus, it is possible to propose a model that structures this specific sentiment analysis application in order to a clearer understanding of the complete process. The model is presented in Figure 5. Since the goal of this paper is to identify the most suitable algorithms to perform the sentiment analysis classification, the data pre-processing and labeling steps as well as the classifiers comparison and validation steps won't be detailed.

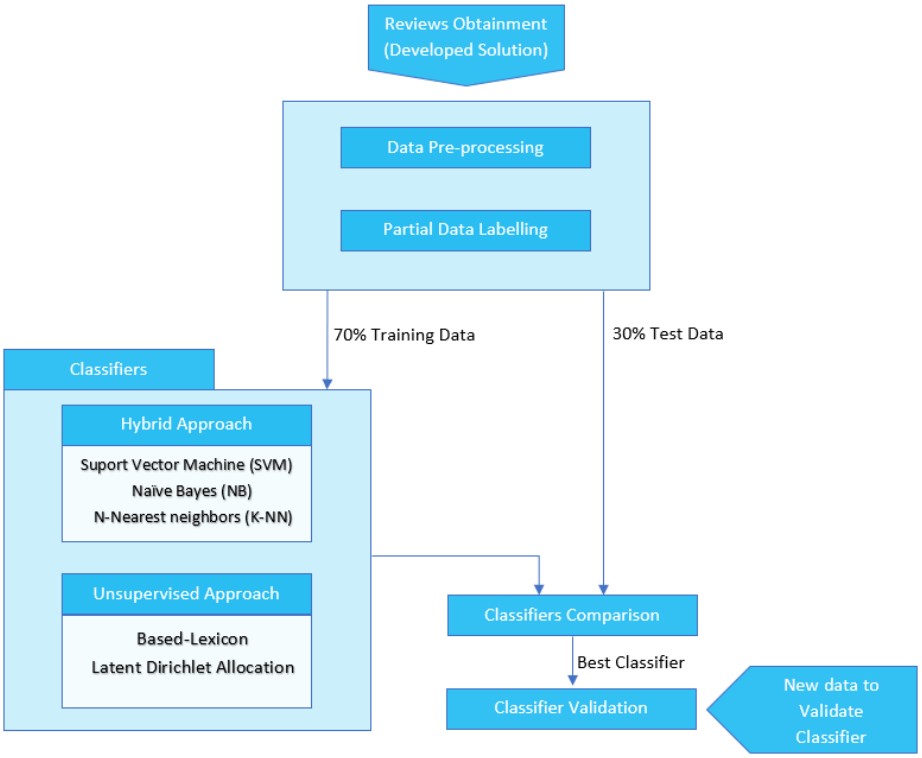

Figure 5 - Sentiment analysis proposed model.

5. CONCLUSION

Even though researches in sentiment analysis became a lot more frequent and accurate, it still is very hard to identify which method, algorithm or strategy should we adopt in order to perform some sort of sentimentl analysis. Probably due to inconsistent results obtained under similar circumstances or even because of the specifications of each project that don't allow us to mimic previous approaches. Either way, it is important to have a global perspective of what could be useful to our project in order to simplify it.

Although the results obtained in this research seemed to be the one that would fit the most to this project, nothing can ensure us that those will be the best ones. Factors like the level of the data pre-processing that we implement or the quality of the obtained data can seriously affect the performance of the algorithms leading to poor results.

Thus, it would be great if we could minimize these factors in order to surpass them. For example, the adoption of Natural Language Processing (NLP), that will be object of study in future work, could help place similar projects on the same level of data-preprocessing, being possible to implement the same methods and have similar results which would seriously improve the method selection process.

ACKNOWLEDGMENTS

This work has been supported by IViSSEM: POCI-01-0145-FEDER-28284, COMPETE: POCI-01-0145-FEDER-007043 and FCT - Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013.

REFERENCES

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages. *ACM Transactions on Information Systems*, *26*(3), 1–34. https://doi.org/10.1145/1361684.1361685

Afzaal, M., Usman, M., Fong, A. C. M., & Fong, S. (2019). Multiaspect-based opinion classification model for tourist reviews. *Expert Systems*, e12371. https://doi.org/10.1111/exsy.12371

Al-amrani, Y., Lazaar, M., Eddine, K., & Kadiri, E. L. (2018). Sentiment Analysis Using Hybrid Method of. *Journal of Theoretical and Applied Information Technology*, *96*(7), 1886–1895. Retrieved from www.jatit.org

Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, *77*, 236–246. https://doi.org/10.1016/j.eswa.2017.02.002

Baek, H., Ahn, J., & Choi, Y. (2012). Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues. *International Journal of Electronic Commerce*, *17*(2), 99–126. https://doi.org/10.2753/jec1086-4415170204

Boudad, N., Faizi, R., Oulad Haj Thami, R., & Chiheb, R. (2018). Sentiment analysis in Arabic: A review of the literature. *Ain Shams Engineering Journal*, *9*(4), 2479–2490. https://doi.org/10.1016/j.asej.2017.04.007

Choi, D., Ko, B., Kim, H., & Kim, P. (2014). Text analysis for detecting terrorism-related articles on the web. *Journal of Network and Computer Applications*, *38*(1), 16–21. https://doi.org/10.1016/j.jnca.2013.05.007

Do, H. H., Prasad, P. W. C., Maag, A., & Alsadoon, A. (2019). Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. *Expert Systems with Applications*, *118*, 272–299. https://doi.org/10.1016/j.eswa.2018.10.003

Erdt, M., Nagarajan, A., Sin, S. C. J., & Theng, Y. L. (2016). Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, *109*(2), 1117–1166. https://doi.org/10.1007/s11192-016-2077-0

Fernández-Gavilanes, M., Àlvarez-López, T., Juncal-Martínez, J., Costa-Montenegro, E., & González-Castaño, F. J. (2015). GTI: An Unsupervised Approach for Sentiment Analysis in Twitter (pp. 533–538). https://doi.org/10.18653/v1/s15-2089

Fortuna, P., & Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, *51*(4), 1–30. https://doi.org/10.1145/3232676

Ganu, G. (2009). Beyond the Stars : Improving Rating Predictions using Review Text Content. *Text*, 1–6. Retrieved from http://www.dbmi.columbia.edu/noemie/ursa

García-Pablos, A., Cuadros, M., & Rigau, G. (2018). W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis. *Expert Systems with Applications*, *91*, 127–137. https://doi.org/10.1016/j.eswa.2017.08.049

Heydari, A., Tavakoli, M. ali, Salim, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems with Applications*, *42*(7), 3634–3642. https://doi.org/10.1016/J.ESWA.2014.12.029

Hoon, L., Vasa, R., Schneider, J.-G., & Mouzakis, K. (2012). A preliminary analysis of vocabulary in mobile app user reviews. In *Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12* (pp. 245–248). New York, New York, USA: ACM Press. https://doi.org/10.1145/2414536.2414578

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04* (p. 168). New York, New York, USA: ACM Press. https://doi.org/10.1145/1014052.1014073

Lei, P., Marfia, G., Pau, G., & Tse, R. (2018). Can we monitor the natural environment analyzing online social network posts? A literature review. *Online Social Networks and Media*, *5*, 51–60. https://doi.org/10.1016/j.osnem.2017.12.001

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review*, *27*, 16–32. https://doi.org/10.1016/j.cosrev.2017.10.002

Mirończuk, M. M., & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, *106*, 36–54. https://doi.org/10.1016/j.eswa.2018.03.058

Moussa, M. E., Mohamed, E. H., & Haggag, M. H. (2018). A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal*, *3*(1), 82–109. https://doi.org/10.1016/j.fcij.2017.12.002

Musto, C., Semeraro, G., & Polignano, M. (2014). A comparison of lexicon-based approaches for sentiment analysis of microblog. In *CEUR Workshop Proceedings* (Vol. 1314, pp. 59–68). Retrieved from http://ceur-ws.org/Vol-1314/paper-06.pdf

Popescu, A. M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural Language Processing and Text Mining* (pp. 9–28). London: Springer London. https://doi.org/10.1007/978-1-84628-754-1_2

Rambocas, M., & Gama, J. (2013). Marketing Research: The Role of Sentiment Analysis. Retrieved from https://pdfs.semanticscholar.org/acd0/c9f75152acd2a622be442d20f96b0a3225d4.pdf

Sabbah, T., Selamat, A., Selamat, M. H., Ibrahim, R., & Fujita, H. (2016). Hybridized term-weighting method for Dark Web classification. *Neurocomputing*, *173*, 1908–1926. https://doi.org/10.1016/j.neucom.2015.09.063

Sammut, C., & Webb, G. I. (Eds.). (2010). *Encyclopedia of Machine Learning*. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-30164-8

Sundermann, C., Domingues, M., Sinoara, R., Marcacini, R., & Rezende , S. (2019). Using Opinion Mining in Context-Aware Recommender Systems: A Systematic Review. *Information*, *10*(2), 42. https://doi.org/10.3390/info10020042

Tavakoli, M., Zhao, L., Heydari, A., & Nenadić, G. (2018). Extracting useful software development information from mobile application reviews: A survey of intelligent mining techniques and tools. *Expert Systems with Applications*, *113*, 186–199. https://doi.org/10.1016/j.eswa.2018.05.037

Ware, M., & Mabe, M. (2015). The STM Report: An overview of scientific and scholarly journal publishing. *Copyright, Fair Use, Scholarly Communication, Etc.*

Webster, J., & Watson, R. T. (2002). Analyzing the Past To Prepare for the Future : Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii.

Zimbra, D., Abbasi, A., Zeng, D., & Chen, H. (2018). The State-of-the-Art in Twitter Sentiment Analysis. *ACM Transactions on Management Information Systems*, *9*(2), 1–29. https://doi.org/10.1145/3185045