A Supervised Approach for Sentiment Analysis using Skipgrams and its Application to Sentiment Visualisation in Social Media

Javier Fernández-Martínez

University of Alicante, Carretera San Vicente del Raspeig S/N, 03690 San Vicente del Raspeig, Alicante, Spain

Abstract

In this Ph.D. thesis we propose, as fundamental research, the design, development and evaluation of a supervised approach for sentiment analysis. This work is based on the hypothesis that an efficient use of the skipgram modelling can improve sentiment analysis tasks and reduce the resources they need. In summary, it consists on a supervised approach that uses machine learning techniques and skipgrams as information units, mainly focused on skipgram selection and filtering. This approach will be evaluated and compared to current state-of-the-art techniques. In addition, as applied research we propose a sentiment visualisation tool, strongly integrated with our sentiment analysis approach. This tool is oriented in the context of social media, measuring reputation and user interactions in real time.

Keywords

sentiment analysis, opinion mining, skipgrams, sentiment visualisation, social monitoring

1. Introduction

Since the creation of Web 2.0, users have become much more active on the Internet, not only by creating new content, but also by commenting on other people's content. This is the reason why we can find a large amount of subjective information on a wide range of topics nowadays. This information can be very valuable for individuals, businesses and public organisations. However, the large amount of subjective information and the fact that it is in textual format, makes it very difficult to exploit it in the right way. Thus, the use *Natural Language Processing* (NLP) techniques become necessary. More specifically, *Sentiment Analysis* (SA) is the NLP subtask that deals with the treatment of subjective texts. Much work has been done on fundamental research in SA in recent years, using many different techniques and resources, from sentiment lexicons [1, 2, 3] to deep learning techniques [4, 5].

However, the extraction of sentiment from texts is often not sufficient to be able to exploit the data properly. Specialised tools are needed to facilitate the visualisation and understanding of the information collected and thus make decisions based on that information. Moreover, in some cases it is important to use this information as soon as possible or in real time, so that

Doctoral Symposium on Natural Language Processing from the PLN.net network 2021 (RED2018-102418-T), 19-20 October 2021, Baeza (Jaén), Spain.

javifm@ua.es (J. Fernández-Martínez)

https://javifmz.github.io (J. Fernández-Martínez)

0000-0002-9552-782X (J. Fernández-Martínez)

resource-intensive SA techniques cannot be used. Again, we can find many works that try to find new ways to visualise this subjective information [6, 7].

In this Ph.D. thesis we propose a new approach to detect subjectivity and polarity in texts, but focused on the subsequent visualisation of subjective information in real time. In Section 2 we will look at related work in this area. Section 3 we will describe the proposal in detail. Finally, in Section 4 we will explain the work done and the work still to be carried out.

2. Related Work

Sentiment Analysis is the task that deals with the computational treatment of opinion, sentiment, and subjectivity in text [8]. This field has several subtasks [9], such as *Aspect-based Sentiment Analysis*, *Subjectivity Detection*, *Emotion Detection* or *Polarity Classification*, but in this work we will focus on the latter. *Polarity Classification* is the task that refers to the classification of an opinionated document as expressing a positive or negative opinion [10]. The approaches that can be followed in this context are usually divided into two main groups [11, 12, 1], *lexicon-based* approaches and *machine-learning-based* approaches.

In *lexicon-based* approaches, the polarity for a document is calculated from the semantic orientation of its words or phrases [13]. These techniques mainly focus on using or building dictionaries of sentiment words. Dictionaries can be created manually [14] or automatically [13]. Examples of general and publicly available sentiment dictionaries include *WordNet Affect* [15], *SentiWordNet* [16] or *ML Senticon* [3]. However, it is difficult to compile and maintain a universal lexicon, as the same word in different domains can express different opinions [13, 17]

The second approach uses *machine learning* techniques. These techniques require the use of a polarity labelled corpus to create a classifier capable of classifying the polarity of new documents. Most of the existing work employs *Support Vector Machines* [18, 19, 20] or *Näıve Bayes* [21, 19, 22], but recent work makes use of *Deep Learning* [4, 23]. In this approach, texts are represented as feature vectors, and a good selection of these features is what mainly improves the performance. These approaches perform very well in the domain in which they have been trained but get worse when used in a different domain [8, 24].

Traditionally, these approaches usually do not take into account the sequentiality of the words contained in the text, so they lose some information during the process. Some techniques can help to solve this problem, such as Transformers or RNNs [25, 26]. The skipgram modelling has also shown good results if used efficiently [27, 28, 29].

Once texts are categorised according to their polarity, it may be necessary to represent in some way the opinion represented in those texts. *Sentiment Summarisation* systems deal with this problem, processing all those opinionated documents and generating a summary that represents the average opinion of all the documents and important aspects of the target addressed in those documents [30]. While these systems usually generate a textual summary, *Sentiment Visualisation* systems do the same but in a graphic or visual manner. *Sentiment Visualisation* is the task that deals with the visualisation and analysis of sentiments discovered in textual data [6]. Techniques in this area have evolved from basic charts used to summarize customer reviews to visual analytics systems dealing with multidimensional datasets, including temporal, relational and geospatial data [6, 7].

3. Proposal

Our proposal consists of a *hybrid approach*, both *lexicon-based* and *machine-learning-based*. A sentiment lexicon is automatically generated from a labelled dataset, assigning a polarity score for a set of selected terms in that dataset, and a machine learning model is responsible for learning how to calculate the polarity of a document based on the terms that appear in the dictionary. The novelty of this work resides on the selection, scoring and filtering of the terms for the dictionary. This work is based on the hypothesis that an efficient use of the skipgram modelling can not only improve sentiment analysis tasks but also reduce the resources needed.

The terms used are *words*, *n-grams* but also *skipgrams*. The *skipgram modelling* technique consists of obtaining n-grams from the words in the text, but allowing some words to be skipped. More specifically, in a *k-skip-n-gram*, *n* determines the number of words, and k the maximum number of words that can be skipped. Skipgrams are thus new terms that retain some of the sequentiality of the original words, but in a more flexible way than n-grams. It is worth noting that an n-gram can be defined as a skipgram where k = 0.

The number of skipgrams generated is usually very large, so that a scoring and filtering process is necessary. The scoring is made taking into account different factors: (i) the number of times the term appears in the corpus; (ii) the number of times the term appears in the corpus for each polarity; (iii) the number of words that the term contains; and (iv) the (average) number of skips required to obtain that term.

The calculation of the polarity of a document is performed by machine learning, where each polarity is considered as a category and each text in the corpus as a learning example. We have followed two different strategies to choose the features of the automatic learning algorithm. The first one is the classic *text classification* approach, using the terms themselves as features for the learning model and the scores in the dictionary as weights. The second strategy performs a combination of the scores generating new features for building the machine learning model.

As applied research we also propose a sentiment visualisation tool, which aims to show as much subjective information as possible at a glance (multiple dimensions) and be useful to help people make decisions based on the data. This tool is oriented in the context of social media, particularly on the social network Twitter[1], measuring the reputation of different entities chosen, and the user interactions related to those entities in real time.

Among all the types of visualisation that we analyse, we can highlight two of them: *evolution of reputation* and *evolution of interactions*. The *evolution of reputation* attempts to give a numerical value to a set of documents according to the size of this set and the polarity detected in its documents, only for a specific points in time, to see how this value has increased or decreased over time. The *evolution of interactions* is similar to the previous one but also focuses on conversations between users of a social network, trying to show how these interactions evolve over time.

[1] https://twitter.com

4. Methodology and future work

Most of the work on this thesis has already been done:

- The polarity classification approach using skipgrams has been developed.
- This approach has been evaluated and compared with some of the existing techniques, carrying out the appropriate experiments in different contexts and different datasets, and multiple articles have been published confirming the effectiveness.
- Different versions of the visualisation tool have been developed over the years, integrating the proposed polarity classification approach.
- Some of these versions has been commercialised and successfully used by many users.

However, there is still some work that we would like to do in order to complete this work:

- Compare the proposal with the state-of-the-art techniques, such as *Deep Learning* and *Word Embeddings*, and integrate them into our approach to improve its effectiveness.
- Compare the performance of skipgrams with respect to the use of other techniques that maintain language sequentiality in the learning process, such as *Transformers* or *RNNs*.
- Perform an appropriate evaluation of the visualisation tool with current similar tools. As it is a visualisation tool, this evaluation should also involve the users of the tool and their experience after use.
- Perform a comparison of the visualisation tool with current similar tools, trying to find possible improvements.
- Publish one or more journal articles with the results obtained.
- Publish and share the code with the research community.

Acknowledgments

This research work has been partially funded by Generalitat Valenciana through project *"SIIA: Tecnologías del lenguaje humano para una sociedad inclusiva, igualitaria, y accesible"* with grant reference PROMETEU/2018/089, and by the Spanish Government and FEDER through the project RTI2018-094653-B-C22: *"Modelang: Modeling the behavior of digital entities by Human Language Technologies"* (*"LIVING-LANG: Living Digital Entities by Human Language Technologies"*).

References

[1] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2011) 267–307.

[2] S. Baccianella, A. Esuli, F. Sebastiani, Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining., in: Lrec, volume 10, 2010, pp. 2200–2204.

[3] F. L. Cruz, J. A. Troyano, B. Pontes, F. J. Ortega, Ml-senticon: Un lexicón multilingüe de polaridades semánticas a nivel de lemas, Procesamiento del Lenguaje Natural 53 (2014) 113–120.

[4] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (2018) e1253.
[5] D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, M. Zhou, Sentiment embeddings with applications to sentiment analysis, IEEE transactions on knowledge and data Engineering 28 (2015) 496–509.
[6] K. Kucher, C. Paradis, A. Kerren, The state of the art in sentiment visualization, in: Computer Graphics Forum, volume 37, Wiley Online Library, 2018, pp. 71–96.
[7] D. Cernea, A. Kerren, A survey of technologies on the rise for emotion-enhanced interaction, Journal of Visual Languages & Computing 31 (2015) 70–86.
[8] B. Pang, L. Lee, Opinion mining and sentiment analysis, Computational Linguistics 35 (2009) 311–312.
[9] L. Yue, W. Chen, X. Li, W. Zuo, M. Yin, A survey of sentiment analysis in social media, Knowledge and Information Systems 60 (2019) 617–663.
[10] A. Montejo-Ráez, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, A knowledge-based approach for polarity classification in twitter, Journal of the Association for Information Science and Technology 65 (2014) 414–425.
[11] M. Annett, G. Kondrak, A comparison of sentiment analysis techniques: Polarizing movie blogs, in: Conference of the Canadian Society for Computational Studies of Intelligence, Springer, 2008, pp. 25–35.
[12] B. Liu, et al., Sentiment analysis and subjectivity., Handbook of natural language processing 2 (2010) 627–666.
[13] P. D. Turney, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, arXiv preprint cs/0212032 (2002).
[14] P. J. Stone, D. C. Dunphy, M. S. Smith, The general inquirer: A computer approach to content analysis. (1966).
[15] C. Strapparava, A. Valitutti, et al., Wordnet affect: an affective extension of wordnet., in: Lrec, volume 4, Lisbon, 2004, p. 40.
[16] F. Sebastiani, A. Esuli, Sentiwordnet: A publicly available lexical resource for opinion mining, in: Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006, pp. 417–422.
[17] G. Qiu, B. Liu, J. Bu, C. Chen, Expanding domain sentiment lexicon through double propagation, in: Twenty-First International Joint Conference on Artificial Intelligence, 2009.
[18] T. Mullen, N. Collier, Sentiment analysis using support vector machines with diverse information sources, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 412–418.
[19] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, Journal of Informetrics 3 (2009) 143–157.
[20] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, S. Patwardhan, Opinionfinder: A system for subjectivity analysis, in: Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, 2005, pp. 34–35.
[21] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, arXiv preprint cs/0409058 (2004).
[22] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in

language, Language resources and evaluation 39 (2005) 165–210.
[23] Q. T. Ain, M. Ali, A. Riaz, A. Noureen, M. Kamran, B. Hayat, A. Rehman, Sentiment analysis using deep learning techniques: a review, Int J Adv Comput Sci Appl 8 (2017) 424.
[24] S. Tan, X. Cheng, Y. Wang, H. Xu, Adapting naive bayes to domain adaptation for sentiment analysis, in: European Conference on Information Retrieval, Springer, 2009, pp. 337–349.
[25] X. Wang, W. Jiang, Z. Luo, Combination of convolutional and recurrent neural network for sentiment analysis of short texts, in: Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers, 2016, pp. 2428–2437.
[26] M. Munikar, S. Shakya, A. Shrestha, Fine-grained sentiment classification using bert, in: 2019 Artificial Intelligence for Transforming Business and Society (AITB), volume 1, IEEE, 2019, pp. 1–5.
[27] Y. Gutierrez, D. Tomas, J. Fernandez, Benefits of using ranking skip-gram techniques for opinion mining approaches, in: eChallenges e-2015 Conference, IEEE, 2015, pp. 1–10.
[28] E. Martınez-Cámara, Y. Gutiérrez-Vázquez, J. Fernández, A. Montejo-Ráez, R. Munoz-Guillena, Ensemble classifier for twitter sentiment analysis, NLP Applications: completing the puzzle (2015) 1–12.
[29] J. Fernández, Y. Gutiérrez, J. M. Gómez, P. Martinez-Barco, Gplsi: Supervised sentiment analysis in twitter using skipgrams, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 2014, pp. 294–299.
[30] S.-A. Bahrainian, A. Dengel, Sentiment analysis and summarization of twitter data, in: 2013 IEEE 16th International Conference on Computational Science and Engineering, IEEE, 2013, pp. 227–234.