

A Supervised Approach for Sentiment Lexicon Generation using Word Skipgrams

Javi Fernández

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

Abstract

This Ph.D. thesis work proposes the design, development and evaluation of a supervised approach for sentiment lexicon generation. It is based on the hypothesis that an efficient use of the skipgram modelling can improve sentiment analysis tasks and reduce the resources needed maintaining an acceptable level of quality. In summary, the novelty of this approach lies in the use of skipgrams as information units and the way they are efficiently generated, weighed and filtered, taking advantage of the useful information they provide about the sequentiality of the language.

Keywords

skipgrams, skipgram modelling, lexicon generation, sentiment lexicon, sentiment analysis

1. Justification for the Proposed Research

The birth of web 2.0 has allowed users to be the main generators of content. Since then, the amount of information has been growing considerably every year, usually unstructured and written in natural language. This information can be turned into knowledge, which makes it very valuable for individuals, businesses and public organisations, but also for scientists with research purposes. However, as the amount of information is large and it is usually in textual format, it is very difficult to exploit it in the right way. Thus, *Natural Language Processing* (NLP) techniques have been essential and have improved substantially over the years.

Part of this information is subjective, that is, it captures the opinions of the users about products, people or other topics, and in some contexts it can be even more valuable. The NLP task that deals with subjective information is *Sentiment Analysis* (SA), and 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 analysis, 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 resource-intensive SA techniques cannot be used.

Doctoral Symposium on Natural Language Processing from the PLN.net network 2022 (RED2018-102418-T), 21-23 September 2022, A Coruña, Spain.


✉ javifm@ua.es (J. Fernández)

🌐 <https://javifmz.github.io> (J. Fernández)

🆔 0000-0002-9552-782X (J. Fernández)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This Ph.D. thesis proposes the design, development and evaluation of a sentiment lexicon generation approach, described in Section 3. Subsequently, in Section 4 we explain the methodology being used to carry out this work. Finally, Section 5 will explain the specific issues of research to be discussed. In the following Section 2 we will look at related work in this area.

2. Background and Related Work

Sentiment Analysis is the task that deals with the computational treatment of opinion, sentiment, and subjectivity in text [6]. This field has several subtasks [7], 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 [8]. The approaches that can be followed in this context are usually divided into two main groups [9, 10, 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 [11]. These techniques mainly focus on using or building dictionaries of sentiment words. Dictionaries can be created manually [12] or automatically [11]. Examples of general and publicly available sentiment dictionaries include *WordNet Affect* [13], *SentiWordNet* [14] 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 [11, 15]

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* [16, 17, 18] or *Näive Bayes* [19, 17, 20], but recent work makes use of *Deep Learning* [4, 21]. 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 [6, 22].

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 [23, 24]. The skipgram modelling has also shown good results if used efficiently [25, 26, 27], and it is the base of our fundamental research.

3. Description of the Proposed Research

The novelty of our research consists on the use of the *skipgram modelling* technique [28] to generate, weight and filter multi-word terms to generate a *sentiment lexicon*, which will be used as the basis for an automatic polarity classifier. The hypothesis of this work is that an efficient use of the skipgram modelling technique can not only improve sentiment analysis tasks but also reduce the resources needed maintaining an acceptable level of quality.

The *skipgram modelling* technique consists of obtaining multi-word terms from a text, similar to n-grams 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. In

this way we are generating additional 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 n-grams can be defined as skipgrams where $k = 0$ (no skips). The skipgram modelling technique is not new in the field of NLP. There are many approaches that use skipgram modelling to relate words to each other, but most of them still use words as the basic unit of information [29, 30].

The main disadvantage of this technique lies in the fact that the number of skipgrams generated is usually very large. To mitigate this problem, a scoring and filtering process becomes necessary. In this work, the scoring and filtering 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.

Furthermore, the weighing and filtering process is not carried out after the generation of all terms, but is done progressively at build time. In a first phase, single-word terms ($n = 1$) are obtained, which are weighted and filtered. From these filtered terms (and only from these), two-word terms ($n = 2$) are obtained, which are also weighted and filtered. The iterative process continues until the desired maximum number of words per term is reached. In this way we manage to create multi-word terms but in a much more efficient way than generating all the skipgrams and filtering them in a last step.

4. Methodology

Most of the work on this research has already been done. The approach has been developed, evaluated and compared to some of the existing techniques, carrying out the appropriate experiments in different contexts and different datasets, and multiple articles have been published confirming its effectiveness [31, 27, 32, 25, 26, 33]. We can highlight the latest work where we have used one of these tools to successfully extract sentiment analysis patterns that determine the virality of tweets about the COVID-19 pandemic [33]. In addition, multiple tools that use this approach (or a previous version) have been developed in the context of this thesis [34, 35, 36, 37]. Some of these tools have been commercialised and successfully used by many clients and businesses.

However, there is still some work to be done. Since this thesis has been extended over time, it is necessary to update the state-of-the-art with the latest similar techniques. In addition, we believe it would be convenient to make an exhaustive study comparing the use of skipgrams with simple words or n-grams in different domains, contexts, textual genres, languages and learning techniques. Moreover, we also plan to do a detailed study on which words work best when creating terms using the skipgram modelling, such as its part of speech, its meaning or its role in the sentence. Finally, we would like to integrate the use of skipgrams into current techniques such as *Word Embeddings* or *Transformers* to see if further improvements can be achieved.

5. Specific Issues of Research to be Discussed

The main questions we want to answer with this thesis are the following:

- Is there any improvement by using skipgrams instead of simple words or n-grams?
- Is it possible to reduce the number of multi-word terms generated by skipgram modelling to improve speed and resource requirements?
- Is the effectiveness in the use of skipgram modelling dependent on the domain, context, textual genre, language or learning technique in which it is applied?
- What kind of words work best using skipgram modelling to generate multi-word terms?

Acknowledgments

This research work has been supported by *TRIVIAL (PID2021-122263OB-C22)* funded by MCIN/AEI/10.13039/501100011033 and by “European Union Regional Development Fund (ERDF) A way of making Europe”, by the “European Union NextGenerationEU/PRTR”.

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] B. Pang, L. Lee, Opinion mining and sentiment analysis, *Computational Linguistics* 35 (2009) 311–312.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] B. Liu, et al., Sentiment analysis and subjectivity., *Handbook of natural language processing* 2 (2010) 627–666.

- [11] P. D. Turney, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, arXiv preprint cs/0212032 (2002).
- [12] P. J. Stone, D. C. Dunphy, M. S. Smith, The general inquirer: A computer approach to content analysis. (1966).
- [13] C. Strapparava, A. Valitutti, et al., Wordnet affect: an affective extension of wordnet., in: *Lrec*, volume 4, Lisbon, 2004, p. 40.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, *Journal of Informetrics* 3 (2009) 143–157.
- [18] 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.
- [19] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, arXiv preprint cs/0409058 (2004).
- [20] J. Wiebe, T. Wilson, C. Cardie, Annotating expressions of opinions and emotions in language, *Language resources and evaluation* 39 (2005) 165–210.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] E. Martinez-Cámara, Y. Gutiérrez-Vázquez, J. Fernández, A. Montejo-Ráez, R. Muñoz-Guillena, Ensemble classifier for twitter sentiment analysis, *NLP Applications: completing the puzzle* (2015) 1–12.
- [27] J. Fernández, Y. Gutiérrez, J. M. Gómez, P. Martínez-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.
- [28] D. Guthrie, B. Allison, W. Liu, L. Guthrie, Y. Wilks, A closer look at skip-gram modelling., in: *LREC*, volume 6, Citeseer, 2006, pp. 1222–1225.

- [29] K. W. Church, Word2vec, *Natural Language Engineering* 23 (2017) 155–162.
- [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [31] J. Fernández Martínez, Y. G. Vázquez, J. M. Gómez Soriano, P. Martínez Barco, A. M. Guijarro, R. M. Guillena, Sentiment analysis of spanish tweets using a ranking algorithm and skipgrams, in: *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural: SEPLN 2013*, Sociedad Española para el Procesamiento del Lenguaje Natural, 2013, pp. 133–142.
- [32] J. Fernández, J. M. Gómez, P. Martínez-Barco, A supervised approach for sentiment analysis using skipgrams, in: *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)*, 2014, pp. 30–36.
- [33] E. Saquete, J. Zubcoff, Y. Gutiérrez, P. Martínez-Barco, J. Fernández, Why are some social-media contents more popular than others? opinion and association rules mining applied to virality patterns discovery, *Expert Systems with Applications* 197 (2022) 116676.
- [34] J. Fernández, Y. Gutiérrez, J. M. Gómez, P. Martínez-Barco, Social rankings: análisis visual de sentimientos en redes sociales, *Procesamiento del Lenguaje Natural* (2015) 199–202.
- [35] J. Fernández, F. Llopis, Y. Gutiérrez, P. Martínez-Barco, Á. Díez, Opinion mining in social networks versus electoral polls., in: *RANLP*, 2017, pp. 231–237.
- [36] J. Fernández, F. Llopis, P. Martínez-Barco, Y. Gutiérrez, Á. Díez, Analizando opiniones en las redes sociales, *Procesamiento del Lenguaje Natural* 58 (2017) 141–148.
- [37] I. Moreno, J. Fernández Martínez, Y. Gutiérrez, et al., *Social-univ 2.0: Tecnologías del lenguaje humano, aplicación para la monitorización omnicanal del entorno social de la universidad de alicante* (2019).