Dancing in the Syntax Forest: Fast, Accurate and Explainable Sentiment Analysis with SALSA

Carlos Gómez-Rodríguez[1], Muhammad Imran[1], David Vilares[1], Elena Solera[2] and Olga Kellert[1]

[1]Universidade da Coruña, CITIC. Department of Computer Science and Information Technologies. 15071 A Coruña, Spain
[2]Universidade da Coruña, Technology Transfer Unit (OTRI). 15071 A Coruña, Spain

Abstract

Sentiment analysis is a key technology for companies and institutions to gauge public opinion on products, services or events. However, for large-scale sentiment analysis to be accessible to entities with modest computational resources, it needs to be performed in a resource-efficient way. While some efficient sentiment analysis systems exist, they tend to apply shallow heuristics, which do not take into account syntactic phenomena that can radically change sentiment. Conversely, alternatives that take syntax into account are computationally expensive. The SALSA project, funded by the European Research Council under a Proof-of-Concept Grant, aims to leverage recently-developed fast syntactic parsing techniques to build sentiment analysis systems that are lightweight and efficient, while still providing accuracy and explainability through the explicit use of syntax. We intend our approaches to be the backbone of a working product of interest for SMEs to use in production.

Keywords

Sentiment analysis, opinion mining, syntax, parsing

1. Introduction

We describe the project "efficient Syntactic Analysis for Large-scale Sentiment Analysis (SALSA)", which is currently being developed by members of the LyS research group at the CITIC research center (Universidade da Coruña). The project is funded by the European Research Council (ERC) with a budget of 150 000 €, under the Proof-of-Concept (PoC) grant scheme, with grant agreement number 101100615. The Proof-of-Concept grant program intends to bridge the gap between basic research results obtained in ERC projects and the early phases of their commercialization, by exploring the innovation potential of the work and bringing it closer to market. The SALSA project started on February 2023 and is scheduled to run until July 2024, with the possibility of an extension.

Given the nature of PoC grants, the project has a scientific research component as well as an innovation and technology transfer component. In Sections 2 and 3, we describe the motivation and methodology, respectively, focusing on the scientific part of the project. Section 4 describes the planning and the project team, with a more global focus that includes the transfer parts as well.

2. Motivation

The problem. The Internet and social media are major platforms where people express their views and share experiences about a variety of topics, including products, services, and events. This results in a wealth of information that can be leveraged to understand public perception, pinpoint product strengths and weaknesses, discover and address people's needs, or track political and market trends. This is the goal of *opinion mining* or *sentiment analysis* systems, which can analyze text and extract sentiment information from it. For instance, a company that releases a new smartphone could gather user opinions from social networks by collecting messages mentioning the product. Given each such message, a sentiment analysis system can determine whether it contains a positive, negative or neutral opinion, both towards the phone as a whole or towards specific aspects like camera or battery.

While many approaches to sentiment analysis have been proposed, an extant challenge lies in the absence of systems that can efficiently process opinions while considering the complex structure of language. While efficient sentiment analysis systems exist, like SentiStrength [1], they rely on shallow heuristic methods that count sentiment words but may struggle with sentences where the overall sentiment is shaped by the grammatical structure. For example, SentiStrength produces the exact same result for "This phone is expensive, and

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, June 19-20, 2024, A Coruña, Spain.
carlos.gomez@udc.es (C. Gómez-Rodríguez); m.imran@udc.es (M. Imran); david.vilares@udc.es (D. Vilares); elena.solera@udc.es (E. Solera); o.kellert@udc.es (O. Kellert)
https://www.grupolys.org/~cgomezr (C. Gómez-Rodríguez); https://scholar.google.com/citations?user=cLXCYOcAAAAJ (M. Imran); https://www.grupolys.org/~david.vilares/ (D. Vilares); https://scholar.google.com/citations?user=XLj-Ol4AAAAJ (O. Kellert)
0000-0003-0752-8812 (C. Gómez-Rodríguez); 0000-0002-4124-7929 (M. Imran); 0000-0002-1295-3840 (D. Vilares); 0000-0003-3541-8303 (E. Solera); 0000-0001-8601-8305 (O. Kellert)

it's not at all good" and "This phone is good, and it's not at all expensive": both sentences actually have opposite sentiment polarities, but the heuristics cannot correctly capture the scope of negation and its influence on how the sentiment should be interpreted.

An alternative is to use approaches that take syntactic stucture into account. Syntax can be injected into a sentiment analysis model either explicitly, by using a parser [2, 3], or implicitly, by using pretrained language models [4] which have been shown to internally represent syntax [5]. The advantage of the former approach over the latter is that it can provide explainability, as an explicit parse tree can be used to show how the polarity is obtained, while the latter can be better in terms of raw accuracy. However, both of these approaches have a high computational cost: pretrained models (especially newer large language models) have considerable CPU and memory requirements [6], while syntactic parsers have traditionally been slow [7, 8]. This makes existing syntactically-guided sentiment analysis approaches infeasible for small entities that have a constrained computational budget.

The solution. The SALSA project proposes to leverage a breakthrough achieved by the previous ERC-funded FASTPARSE project [7] to build sentiment analysis systems that are efficient and, at the same time, use explicit syntax to improve accuracy and provide explainability. The FASTPARSE project had the goal to substantially improve the speed of syntactic parsing, and was a resounding success, advancing the state of the art in both parsing speed and accuracy several times, and achieving accurate parsers that improved speed by an order or magnitude over previous approaches.

In particular, within the FASTPARSE project, we defined a new syntactic parsing paradigm that casts parsing as a sequence labeling task by defining encodings that can represent syntax trees by means of discrete labels associated with words. This was shown to work for the two main families of grammatical formalisms, with encodings that proved viable for constituency parsing [9] and dependency parsing [10], and new encodings have kept being defined since then that further push parsing accuracy [11, 12].

For the purposes of this project, the approach to parsing as sequence labeling has two especially relevant advantages:

1. Pluggability: parse trees are represented with tags associated with words, a standard representation that is easy to feed to downstream tasks, be it as features [13] or by using multitask learning [14].
2. Speed: these parsers can parse around a thousand sentences per second on a standard consumer GPU [9, 10, 15].

Our proposal is, thus, to use this new parsing paradigm to create efficient syntax-guided sentiment analysis systems that can operate at a large scale with modest computational resources. This will democratize accurate large-scale sentiment analysis technology, putting it within reach of smaller institutions and companies that cannot afford to deploy slow parsers or language models.

3. Methodology

Our project addresses two variants (or subtasks) of sentiment analysis:

(a) Polarity classification: a coarse-grained task where the input is a text, such as a social network message, and the output is its polarity (i.e. a representation, typically in a discrete scale, of how positive or negative the global opinion expressed in the text is).

(b) Fine-grained sentiment analysis, often called aspect-based sentiment analysis [16]: in this task, the goal is not to obtain a single value for the whole text, but individual opinions about the different targets that may appear in the message (for example, the text "I really like this phone's camera, but the battery life is not acceptable" contains both a highly positive opinion about the camera and a negative one about the battery).

For this purpose, we will explore three approaches to integrate fast syntactic parsing into solving these two sentiment analysis tasks:

- Rule-based pipeline approach: we start by using a fast dependency parser to obtain the syntactic tree for an input sentence. We then use syntax-based rules and a polarity lexicon to navigate the tree, assigning polarity values to its syntactic components (addressing subtask (b) above). These values can then be integrated by the rules to obtain an overall polarity for the whole sentence (addressing subtask (a)). The syntax-based rules are specifically designed to address syntactic phenomena that influence opinion polarity, such as negation (for instance, in the example above, "not acceptable" indicates a negative sentiment because the negation alters the positive meaning of "acceptable"), intensification (the phrase "really like" signifies a stronger preference than merely "like"), and adversatives (where "but" juxtaposes a positive statement against a negative one). It is worth noting that this kind of approach have been applied in the past, particularly for subtask (a) (see [2, 3]), but the use of slow parsers at the time rendered large-scale application costly.

- Multitask learning approach: thanks to the parsing-as-sequence-labeling approach, we can train a multitask learning model to perform sentiment analysis and syntactic parsing simultaneously. This allows each task to benefit from the insights of the other implicitly, without relying on explicit syntax-based rules. For subtask (a), this technique can be implemented by formulating the sentiment analysis sequence labeling task in such a way that the polarity of the whole sentence is represented by a tag associated with the last word. For subtask (b), we will design a sequence labeling encoding to represent fine-grained opinoins.
- Integrated approach: this approach is specific to subtask (b). We represent the sentiment structure of the sentence as a tree. This means that we can apply the same algorithms that we use for syntactic parsing directly to perform sentiment analysis. Furthermore, this strategy can be combined with the multitask learning approach described above, enabling the training of a sequence labeling model to simultaneously produce two types of trees: one for syntactic parsing and another for sentiment analysis.

Each of the described approaches will be evaluated on freely-available sentiment analysis corpora. We will take a multilingual focus, ensuring that our systems work on a diverse range of languages (for the dependency parsing part, this is ensured by the use of Universal Dependencies [17]). For the initial value proposition, we will focus on Spanish and English, since our main target is the local Spanish market. However, the core technology should be available to easily adapt to more languages when needed.

In particular, a good starting point when it comes to datasets are the SemEval 2022 Task 10 corpora [18], as they cover five different languages and their annotation includes individual polarity targets, thus supporting both our subtasks (a) and (b). Given our focus on the Spanish market, we are also working with a corpus of TripAdvisor reviews in Spanish, from Rest-Mex 2023 [19]. This also allows us to evaluate performance in a noisy context. Some preliminary results of the first (rule-based pipeline) approach on the latter corpus can be found in [20].

4. Planning and Team

The SALSA project is organized into three work packages. The first one addresses the scientific research part of the project, and hence it is the one we have addressed in Sections 2 and 3. The other two correspond to the innovation and technology transfer parts of the project. In particular, the work packages are as follows:

- WP1: Technical and Performance Validation. This involves three tasks:
  - Validation of fast parsers for polarity classification: the main objective is to adapt the parsers developed in the FASTPARSE project to work on polarity classification tasks. To achieve this, we have created linguistic rules that, when integrated with a parser's output, allow us to determine the polarity of sentences by analyzing the words and syntactic structures in the text [20]. Additionally, we are exploring a purely machine learning-based method that does not rely on these rules. This method involves training a parser capable of sequence labeling in conjunction with a sequence model that uses discrete tags to represent sentence polarity, employing a multitask learning framework. The success of this task is assessed through the validation of various models' ability to classify polarity, as evidenced by performance metrics.
  - Validation of fast parsers for fine-grained sentiment analysis: in this task, we aim to go beyond polarity classification by employing syntactic parsers for more detailed sentiment analysis. The goal is to analyze sentiments at a more granular level, capturing specific opinions within a sentence. This involves recognizing how a single sentence might simultaneously provide positive feedback on one aspect of a product while criticizing another. To achieve this, we will adapt the approaches based on linguistic rules and on multitask learning to process fine-grained information: the rules will be used to infer sentiment information for smaller linguistic units apart from the whole sentence, and a sequence labeling encoding will be used to represent fine-grained sentiment information for the multitask setup. In addition, we will try a third setup where the fine-grained sentiment information will be represented as a tree, so that syntactic parsers can learn to output it directly.
  - Cross-domain generalization. The performance of sentiment analysis models trained on a given domain (e.g. restaurant reviews) often degrades when they are used in a different domain (e.g. movie reviews). Since our project is targeted primarily towards making the technology ac-

cessible to small entities, which will often lack access to high-quality in-domain corpora, it is important to ensure that our systems can adapt to different domains. In this respect, rule-based approaches have been shown to be better than purely supervised approaches [3]. In this task, we will evaluate the generalization capabilities of the various models developed in the previous two tasks to different domains, and explore specific domain adaptation approaches (like the adaptation of sentiment dictionaries [2]) if needed.

- WP2: Market Research and Validation. This involves three tasks:
  - Market Analysis. We have performed a study to identify current sentiment analysis industry trends, market need and market size at the European level. This study, which was subcontracted to an external consultancy company, includes the potential market niches for our sentiment analysis models, identification of comparative technologies and information on potential user needs.
  - Validation of value proposition. This involves conducting interviews, as well as Design Thinking workshops, to identify target user needs and validate our proposition.
  - Commercial viability: product-market fit. This task consists in involving potential users (identified in the previous tasks) in iterative testing and validation of the developed models.
- WP3: Pre-Technology Transfer. This includes two tasks:
  - Creating SALSA business model canvas. We will refine our business model canvas through a process of continuous improvement. This will involve defining, enhancing, and validating our value proposition, primary revenue streams, and cost structures, alongside identifying the crucial partners, customers, channels, and business relationships that form the core components of an open source business model. We will assess the viability of establishing a spin-off to offer service agreements related to SALSA.
  - Building a business environment. The aim of this task is to establish contact with potential partners and stakeholders that can advance the TRL and scale the technology, as well as with potential future investors. We will also considering the possibility of preparing an EIC Transition proposal, depending on the outcomes of previous tasks.

The project team is composed by the authors of this paper: a postdoctoral researcher (O. Kellert), an MSCA predoctoral researcher (M. Imran), a technical manager (D. Vilares) and a project manager (E. Solera), led by the PI (C. Gómez-Rodríguez). In addition, some of the market-oriented tasks of the project are being carried out with the help of Matical Innovation, which has been subcontracted for that purpose.

Acknowledgments

This project has received funding by the European Research Council (ERC), under the Horizon Europe research and innovation programme (SALSA, grant agreement No 101100615).

References

[1] M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web, J. Am. Soc. Inf. Sci. Technol. 63 (2012) 163–173. URL: https://doi.org/10.1002/asi.21662. doi:10.1002/asi.21662.

[2] D. Vilares, M. A. Alonso, C. Gómez-Rodríguez, A syntactic approach for opinion mining on Spanish reviews, Natural Language Engineering 21 (2015) 139–163.

[3] D. Vilares, C. Gómez-Rodríguez, M. A. Alonso, Universal, unsupervised (rule-based), uncovered sentiment analysis, Knowledge-Based Systems 118 (2017) 45 – 55. URL: http://www.sciencedirect.com/science/article/pii/S0950705116304701. doi:http://dx.doi.org/10.1016/j.knosys.2016.11.014.

[4] M. Hoang, O. A. Bihorac, J. Rouces, Aspect-based sentiment analysis using BERT, in: M. Hartmann, B. Plank (Eds.), Proceedings of the 22nd Nordic Conference on Computational Linguistics, Linköping University Electronic Press, Turku, Finland, 2019, pp. 187–196. URL: https://aclanthology.org/W19-6120.

[5] D. Vilares, M. Strzyz, A. Søgaard, C. Gómez-Rodríguez, Parsing as pretraining, in: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 9114–9121. URL: https://aaai.org/ojs/index.php/

AAAI/article/view/6446. doi:https://doi.org/10.1609/aaai.v34i05.6446.

[6] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://aclanthology.org/P19-1355. doi:10.18653/v1/P19-1355.

[7] C. Gómez-Rodríguez, Towards fast natural language parsing: FASTPARSE ERC Starting Grant, Procesamiento del Lenguaje Natural 59 (2017) 121–124.

[8] C. Gómez-Rodríguez, I. Alonso-Alonso, D. Vilares, How important is syntactic parsing accuracy? An empirical evaluation on rule-based sentiment analysis, Artificial Intelligence Review 52 (2019) 2081–2097. URL: https://doi.org/10.1007/s10462-017-9584-0. doi:10.1007/s10462-017-9584-0.

[9] C. Gómez-Rodríguez, D. Vilares, Constituent parsing as sequence labeling, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2018, pp. 1314–1324. URL: http://aclweb.org/anthology/D18-1162.

[10] M. Strzyz, D. Vilares, C. Gómez-Rodríguez, Viable dependency parsing as sequence labeling, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 717–723. URL: https://www.aclweb.org/anthology/N19-1077.

[11] A. Amini, T. Liu, R. Cotterell, Hexatagging: Projective dependency parsing as tagging, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1453–1464. URL: https://aclanthology.org/2023.acl-short.124. doi:10.18653/v1/2023.acl-short.124.

[12] C. Gómez-Rodríguez, D. Roca, D. Vilares, 4 and 7-bit labeling for projective and non-projective dependency trees, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6375–6384. URL: https://aclanthology.org/2023.emnlp-main.393. doi:10.18653/v1/2023.emnlp-main.393.

[13] Y. Wang, M. Johnson, S. Wan, Y. Sun, W. Wang, How to best use syntax in semantic role labelling, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5338–5343. URL: https://aclanthology.org/P19-1529. doi:10.18653/v1/P19-1529.

[14] M. Strzyz, D. Vilares, C. Gómez-Rodríguez, Sequence labeling parsing by learning across representations, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5350–5357. URL: https://www.aclweb.org/anthology/P19-1531.

[15] M. Anderson, C. Gómez-Rodríguez, A modest Pareto optimisation analysis of dependency parsers in 2021, in: S. Oepen, K. Sagae, R. Tsarfaty, G. Bouma, D. Seddah, D. Zeman (Eds.), Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021), Association for Computational Linguistics, Online, 2021, pp. 119–130. URL: https://aclanthology.org/2021.iwpt-1.12. doi:10.18653/v1/2021.iwpt-1.12.

[16] G. Brauwers, F. Frasincar, A survey on aspect-based sentiment classification, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3503044. doi:10.1145/3503044.

[17] D. Zeman, et al., Universal Dependencies 2.12, 2023. URL: http://hdl.handle.net/11234/1-5150, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[18] J. Barnes, L. Oberlaender, E. Troiano, A. Kutuzov, J. Buchmann, R. Agerri, L. Øvrelid, E. Velldal, SemEval 2022 task 10: Structured sentiment analysis, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 1280–1295. URL: https://aclanthology.org/2022.semeval-1.180. doi:10.18653/v1/2022.semeval-1.180.

[19] M. Ángel Álvarez-Carmona y Ángel Díaz-Pacheco y Ramón Aranda y Ansel Yoan Rodríguez-González y Victor Muñiz-Sánchez y Adrián Pastor López-Monroy y Fernando Sánchez-Vega y Lázaro Bustio-Martínez, Overview of rest-mex at iberlef 2023: Research on sentiment analysis task for mexican tourist texts, Procesamiento del Lenguaje Natural 71 (2023) 425–436. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6572.

[20] O. Kellert, M. U. Zaman, N. H. Matlis, C. Gómez-Rodríguez, Experimenting with ud adaptation of an unsupervised rule-based approach for sentiment analysis of Mexican tourist texts, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), volume 3946 of *CEUR Workshop Proceedings*, Jaén, Spain, 2023. URL: https://ceur-ws.org/Vol-3496/restmex-paper15.pdf.