

On the use of sentiment analysis for linguistics research. Observations on sentiment polarity and the use of the progressive in Italian

Lorella Viola^{a*}

^aLuxembourg Centre for Contemporary and Digital History (C²DH), University of Luxembourg, Luxembourg

Belval Campus - Maison des Sciences Humaines 11, Porte des Sciences L-4366, Esch-sur-Alzette, Email: lorella.viola@uni.lu, Twitter: @ViolaLorella, Personal website: <https://lorellaviola.me.uk>, OrcId: 0000-0001-9994-0841

Introduction

Today computational methods are used more and more in domains outside of their original conception. For example, sentiment analysis (SA) – a technique originally designed to infer general opinions from product reviews – is applied to make stock market predictions, in the health sector, by government agencies and in academic research (e.g., social science). In line with one of the themes of this year's DH Benelux conference, this presentation discusses the challenges of integrating quantitative, statistical and computational methods and techniques into the Arts and Humanities. Specifically, it investigates the suitability of SA for linguistics research, where it has not been explored yet. Indeed, whereas this technique has been used across disciplines as a means to answer a variety of research questions, such as to analyse opinions about social and political issues, particularly on social media (see Salas-Zárate et al. 2017; X. Liu 2022; Zhong and Ren 2022, Haselmayer and Jenny 2017; Ansari et al. 2020; Matalon et al. 2021, Karamibekr and Ghorbani 2012; Nguyen et al. 2020; Bhat et al. 2020, Thelwall 2016; Burscher, Vliegenthart, and Vreese 2016; Amarasekara and Grant 2019) in linguistics research, the use of SA has largely been overlooked. The aim is to investigate the limitations and advantages of this method, its applicability for tasks outside of its original conception, and to what extent it may be used for linguistics research.

In doing so, this study responds to recent calls for new analytical methods that could expand the field of linguistics by devising novel and dynamic ways of exploring established topics (Rose and McKinley 2020). Specifically, I test SA against an established case in linguistics: the correlation between subjective attitudes such as surprise, irritation and discontent and the use of the

progressive. Several accounts in various languages (e.g., English, German, Dutch) (Traugott and Dasher 2001; Killie 2004; Levin 2013; Pfaff, Bergs, and Hoffmann 2013; Anthonissen, De Wit, and Mortelmans 2016; 2019) have indeed observed that non-aspectual pragmatic or subjective meanings, for example to signal politeness and discontent, may mark the use of progressive constructions. The use case is the Italian language. In this way, the study also provides fresh insights into this line of enquiry in which, with the exception of limited work on French (e.g., De Wit, Petré, and Brisard 2020), works have been largely focused on Germanic languages and predominantly on English. Furthermore, in this way this contribution aligns with the general theme of this year's conference "Crossing borders: digital humanities research across languages and modalities".

Analysis and results

For the analysis, the study uses FEEL-IT (Bianchi, Nozza, and Hovy 2021), a state-of-the-art transformer-based machine learning classifier for emotion and sentiment classification in Italian to predict sentiments and identify the corresponding emotions. It employs the Italian BERT model UmBERTo trained on Commoncrawl ITA (Parisi, Francia, and Magnani [2020] 2021) and fine-tuned on a corpus by the same name containing Italian Twitter posts manually annotated with four basic emotions: anger, fear, joy, sadness. This SA classifier returns a prediction of the sentiment polarity (i.e., positive or negative) and the specific corresponding sentiment (i.e., anger, joy, sadness, fear). The sentiment polarity is obtained by collapsing the four emotions into the two categories (i.e., joy \square positive; fear, anger, sadness \square negative). Another challenge of using computational methods for linguistic research is related to the predominance of models, datasets, and tools devised and developed for the English language. Even when models and resources in other languages exist, they are often proprietary and expensive (e.g., Google Cloud Platform Console) as well as not transparent and often offering only opaque documentation. Because not all languages are equally resourced digitally and computationally, researchers, teachers and curators are forced to compromise on which tasks can be performed, with which tools and through which platforms (Viola 2022b). Such Anglophone-centricity is therefore often still a barrier for researchers working with languages other than English. Resources like FEEL-IT are important to counterbalance the predominance of English in computer science. The dataset for the analysis is the Italian language data used in the EVALITA Parsing Task project. This is a periodic evaluation campaign of Natural Language Processing (NLP) and speech tools for the Italian language. The

aim of this project is to promote the development of language and speech technologies for the Italian language and to offer to the scientific community a shared framework where different systems and approaches can be evaluated consistently. The corpus used here is the training and test data corpora used for the 2014 EVALITA task. It contains language samples from various sources (e.g., legal texts, news articles, Wikipedia articles) collected from 2011 to 2014 and it totals up to 238,556 words. After launching the SA model on the dataset, random excerpts are analysed qualitatively to assess whether the assigned prediction can be reputed reliable.

The language material was first tokenised into sentences (13,343 sentences) and scanned for verbs at the progressive form. This preliminary investigation filtered the dataset only for relevant data, that is to work on a subset containing gerund and progressive. The obtained subset included 628 sentences containing 377 verbs at the gerund/progressive form of which 382 from the first group and 248 from the second (none from the third). The subset was then queried for sentiment using the FEEL-IT Python library. The results showed that the majority of the sentences in the subset (469 vs 159) were classified as negative. Following the quantitative analysis, the results were tested for significance (Figure 1); The Chi-square results show statistically significant correlations between negative sentiment and the use of gerund/progressive constructions thus aligning with previous findings in other languages. At the same time Cramer's phi shows that, though the correlation is significant, the strength of the relationship is not high. This finding could be explained by several factors (e.g., diversity of sources, different genres, short sentences, very limited context provided). Future investigations could for example perform logistic regression to identify other potential factors not measured in this study that are predictive of sentiment.

Finally, qualitative analysis of random excerpts assesses to what extent the sentiment and emotion classification can be reputed reliable. The results of this analysis will be discussed during the presentation; they demonstrate that the technique is far from being methodologically ideal, particularly for empirical social research and it highlighted some of the uncertainties of using this technique in fields other than IR and for tasks other than product review. For example, it was found a high degree of discrepancy between the writer's intention, the way the classification was carried out for training the SA model, and the text to which the model was applied. It is argued here that these uncertainties add further complexity and ambiguity to the already existing limitations of the technique which should be carefully assessed by researchers and practitioners before and during the analysis.

Conclusions

Traditional methods such as words' collocation analyses for example, may no longer be sufficient for identifying otherwise not immediately evident patterns and discontinuities. More sophisticated methods such as SA can therefore be of great assistance to linguists who now are more and more confronted with the challenge of analysing the complexities of the textual material produced and easily available. However, although undoubtedly still providing powerful means to navigate large quantities of language material, these methods should not be adopted uncritically and researchers and practitioners should always resize their epistemological expectations accordingly.

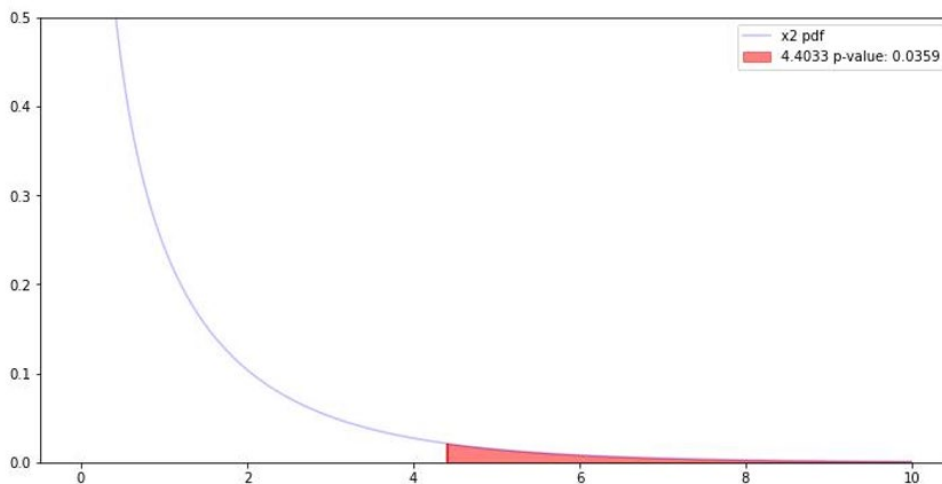


Figure 1: Visualisation of Chi-square results. The red section represents the p-value.

References

- Amarasekara, Inoka, and Will J Grant. 2019. 'Exploring the YouTube Science Communication Gender Gap: A Sentiment Analysis'. *Public Understanding of Science* 28 (1): 68–84. <https://doi.org/10.1177/0963662518786654>.
- Ansari, Mohd Zeeshan, M.B. Aziz, M.O. Siddiqui, H. Mehra, and K.P. Singh. 2020. 'Analysis of Political Sentiment Orientations on Twitter'. *Procedia Computer Science* 167: 1821–28. <https://doi.org/10.1016/j.procs.2020.03.201>.
- Anthonissen, Lynn, Astrid De Wit, and Tanja Mortelmans. 2016. 'Aspect Meets Modality: A Semantic Analysis of the German *Am* -Progressive'. *Journal of Germanic Linguistics* 28 (1): 1–30. <https://doi.org/10.1017/S1470542715000185>.
- . 2019. '(Inter)Subjective Uses of the Dutch Progressive Constructions'. *Linguistics* 57 (5): 1111–59. <https://doi.org/10.1515/ling-2019-0019>.

- Bhat, Muzafar, Monisa Qadri, Noor-ul-Asrar Beg, Majid Kundroo, Naffi Ahanger, and Basant Agarwal. 2020. 'Sentiment Analysis of Social Media Response on the Covid19 Outbreak'. *Brain, Behavior, and Immunity* 87 (July): 136–37. <https://doi.org/10.1016/j.bbi.2020.05.006>.
- Bianchi, Federico, Debora Nozza, and Dirk Hovy. 2021. 'FEEL-IT: Emotion and Sentiment Classification for the Italian Language'. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 76–83. Online: Association for Computational Linguistics. <https://www.aclweb.org/anthology/2021.wassa-1.8>.
- Bland, Susan Kesner. 1988. 'The Present Progressive in Discourse: Grammar versus Usage Revisited'. *TESOL Quarterly* 22 (1): 53. <https://doi.org/10.2307/3587061>.
- Bobicev, Victoria, and Marina Sokolova. 2017. 'Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective'. In *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, 97–102. Incoma Ltd. Shoumen, Bulgaria. https://doi.org/10.26615/978-954-452-049-6_015.
- . 2018. 'Thumbs Up and Down: Sentiment Analysis of Medical Online Forums'. In *EMNLP 2018*. <https://doi.org/10.18653/v1/W18-5906>.
- Bosco, Cristina, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. 'The EVALITA 2014 Dependency Parsing Task'. In *EVALITA 2014 Evaluation of NLP and Speech Tools for Italian*, 1–8. Pisa University Press.
- Burscher, Bjorn, Rens Vliegthart, and Claes H. de Vreese. 2016. 'Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue'. *Social Science Computer Review* 34 (5): 530–45. <https://doi.org/10.1177/0894439315596385>.
- De Wit, Astrid, and Adeline Patard. 2013. 'Modality, Aspect and the Progressive: The Semantics of the Present Progressive in French in Comparison with English'. *Languages in Contrast* 13 (1): 113–32. <https://doi.org/10.1075/lic.13.1.06wit>.
- De Wit, Astrid, Peter Petré, and Frank Brisard. 2020. 'Standing out with the Progressive'. *Journal of Linguistics* 56 (3): 479–514. <https://doi.org/10.1017/S0022226719000501>.
- Freund, Nina. 2016. 'Recent Change in the Use of Stative Verbs in the Progressive Form in British English: I'm Loving It'. *Language Studies Working Papers* 7: 50–61.
- González-Bailón, Sandra, and Georgios Paltoglou. 2015. 'Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources'. *The ANNALS of the American Academy of Political and Social Science* 659 (1): 95–107. <https://doi.org/10.1177/0002716215569192>.
- Haselmayer, Martin, and Marcelo Jenny. 2017. 'Sentiment Analysis of Political Communication: Combining a Dictionary Approach with Crowdcoding'. *Quality & Quantity* 51 (6): 2623–46. <https://doi.org/10.1007/s11135-016-0412-4>.
- Jini Jojo Stephen, and P. Prabu. 2019. 'Detecting the Magnitude of Depression in Twitter Users Using Sentiment Analysis'. *International Journal of Electrical and Computer Engineering (IJECE)* 9 (4): 3247. <https://doi.org/10.11591/ijece.v9i4.pp3247-3255>.

- Karamibekr, Mostafa, and Ali A. Ghorbani. 2012. 'Sentiment Analysis of Social Issues'. In *2012 International Conference on Social Informatics*, 215–21. <https://doi.org/10.1109/SocialInformatics.2012.49>.
- Karppi, Tero, and Kate Crawford. 2016. 'Social Media, Financial Algorithms and the Hack Crash'. *Theory, Culture & Society* 33 (1): 73–92. <https://doi.org/10.1177/0263276415583139>.
- Kawade, Dipak R., and Kavita S. Oza. 2017. 'Sentiment Analysis: Machine Learning Approach'. *International Journal of Engineering and Technology* 9 (3): 2183–86.
- Killie, Kristin. 2004. 'Subjectivity and the English Progressive'. *English Language and Linguistics* 8 (1): 25–46. <https://doi.org/10.1017/S1360674304001236>.
- Langacker, Ronald W. 1983. *Foundations of Cognitive Grammar*. Indiana University Linguistics Club.
- Lee, Sophia Yat Mei, and Helena Yan Ping Lau. 2020. 'An Event-Comment Social Media Corpus for Implicit Emotion Analysis'. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 1633–42. Marseille, France: European Language Resources Association. <https://www.aclweb.org/anthology/2020.lrec-1.203>.
- Levin, Magnus. 2013. 'The Progressive Verb in Modern American English'. In *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, edited by Bas Aarts, Geoffrey Leech, Joanne Close, and Sean Wallis, 187–216. Studies in English Language. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139060998.009>.
- Liu, Bing. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge university press.
- Liu, Xin. 2022. 'Analysis of Psychological Characteristics and Emotional Expression Based on Deep Learning in Higher Vocational Music Education'. *Frontiers in Psychology* 13 (September): 981738. <https://doi.org/10.3389/fpsyg.2022.981738>.
- Martínez-Vázquez, Montserrat. 2018. 'I'm Loving It! A Corpus-Based Study of the Progress of Love'. *Journal of English Linguistics* 46 (2): 140–66. <https://doi.org/10.1177/0075424218765609>.
- Matalon, Yogev, Ofir Magdaci, Adam Almozlino, and Dan Yamin. 2021. 'Using Sentiment Analysis to Predict Opinion Inversion in Tweets of Political Communication'. *Scientific Reports* 11 (1): 7250. <https://doi.org/10.1038/s41598-021-86510-w>.
- Moreno-Ortiz, Antonio. 2017. 'Lingmotif: Sentiment Analysis for the Digital Humanities'. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 73–76. Valencia, Spain: Association for Computational Linguistics. <https://aclanthology.org/E17-3019>.
- Moreno-Ortiz, Antonio, Javier Fernandez-Cruz, and Chantal Pérez Chantal Hernández. 2020. 'Design and Evaluation of SentiEcon: A Fine-Grained Economic/Financial Sentiment Lexicon from a Corpus of Business News'. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 5065–72. Marseille, France: European Language Resources Association. <https://www.aclweb.org/anthology/2020.lrec-1.623>.
- Munika, Manish, Sushil Shakya, and Aakash Shrestha. 2019. 'Fine-Grained Sentiment Classification Using BERT'. arXiv. <https://doi.org/10.48550/arXiv.1910.03474>.

- Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. 'How We Do Things With Words: Analyzing Text as Social and Cultural Data'. *Frontiers in Artificial Intelligence* 3. <https://www.frontiersin.org/articles/10.3389/frai.2020.00062>.
- Pang, Bo, and Lillian Lee. 2008. 'Opinion Mining and Sentiment Analysis'. *Foundations and Trends® in Information Retrieval* 2 (1–2): 1–135. <https://doi.org/10.1561/1500000011>.
- Parisi, Loreto, Simone Francia, and Paolo Magnani. (2020) 2021. 'Musixmatchresearch/Umberto'. Python. Musixmatch Research. <https://github.com/musixmatchresearch/umberto>.
- Pfaff, Meike, Alexander Bergs, and Thomas Hoffmann. 2013. 'I Was Just Reading This Article— on the Expression of Recentness and the English Past Progressive'. In *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, edited by Bas Aarts, Geoffrey Leech, Joanne Close, and Sean Wallis, 217–38. Studies in English Language. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9781139060998.010>.
- Puschmann, Cornelius, and Alison Powell. 2018. 'Turning Words Into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media'. *Social Media + Society* 4 (3): 205630511879772. <https://doi.org/10.1177/2056305118797724>.
- Rose, Heath, and Jim McKinley, eds. 2020. *The Routledge Handbook of Research Methods in Applied Linguistics*. Routledge Handbooks in Applied Linguistics. New York: Routledge.
- Salas-Zárate, María Pilar, Mario Andrés Paredes-Valverde, Miguel Ángel Rodríguez-García, Rafael Valencia-García, and Giner Alor-Hernández. 2017. 'Sentiment Analysis Based on Psychological and Linguistic Features for Spanish Language'. In *Current Trends on Knowledge-Based Systems*, edited by Giner Alor-Hernández and Rafael Valencia-García, 120:73–92. Intelligent Systems Reference Library. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-51905-0_4.
- Scheffer, Johannes. 1975. *The Progressive in English*. Amsterdam : North-Holland Pub. Co. ; New York : American Elsevier Publishing Company. <http://archive.org/details/progressiveineng0000sche>.
- Schmidt, Thomas, Dangel, Johanna, and Wolff, Christian. 2021. 'SentText: A Tool for Lexicon-Based Sentiment Analysis in Digital Humanities'. In . Universität Regensburg. <https://doi.org/10.5283/EPUB.44943>.
- Smitherberg, Erik. 2005. *The Progressive in 19th-Century English: A Process of Integration*. Language and Computers, N° 54. Amsterdam: Rodopi.
- Sprugnoli, Rachele. 2021. 'MultiEmotions-It: A New Dataset for Opinion Polarity and Emotion Analysis for Italian'. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, edited by Johanna Monti, Felice Dell'Orletta, and Fabio Tamburini. Vol. 2769. CEUR Workshop Proceedings. Bologna, Italy: CEUR. <https://ceur-ws.org/Vol-2769/#8>.
- Swan, Toril, and Leiv Egil Breivik. 1997. 'Subject-Oriented Adverbs in a Diachronic and Contrastive Perspective'. In *Language History and Linguistic Modelling*, edited by Raymond Hickey and Stanislav Puppel, 1997th ed., 101:395–422. Trends in Linguistics. Studies and Monographs. Berlin, New York: DE GRUYTER MOUTON. <https://doi.org/10.1515/9783110820751.395>.
- . 2011. 'English Sentence Adverbials in a Discourse and Cognitive Perspective'. *English Studies* 92 (6): 679–92. <https://doi.org/10.1080/0013838X.2011.604917>.

- Thelwall, Mike. 2016. 'Sentiment Analysis'. In *The SAGE Handbook of Social Media Research Methods*, by Luke Sloan and Anabel Quan-Haase, 545–56. 1 Oliver's Yard, 55 City Road London EC1Y 1SP: SAGE Publications Ltd. <https://doi.org/10.4135/9781473983847.n32>.
- Traugott, Elizabeth Closs. 1989. 'On the Rise of Epistemic Meanings in English: An Example of Subjectification in Semantic Change'. *Language* 65 (1): 31–55. <https://doi.org/10.2307/414841>.
- . 1995. 'Subjectification in Grammaticalisation'. In *Subjectivity and Subjectivisation: Linguistic Perspectives*, edited by Dieter Stein and Susan Wright, 31–54. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511554469.003>.
- Traugott, Elizabeth Closs, and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511486500>.
- Viola, Lorella. 2022a. 'Networks of Migrants' Narratives: A Post-Authentic Approach to Heritage Visualisation'. *Journal on Computing and Cultural Heritage*.
- . 2022b. *The Humanities in the Digital: Beyond Critical Digital Humanities*. S.l.: Palgrave Macmillan.
- Zhong, Ni, and JunBao Ren. 2022. 'Using Sentiment Analysis to Study the Relationship between Subjective Expression in Financial Reports and Company Performance'. *Frontiers in Psychology* 13 (July): 949881. <https://doi.org/10.3389/fpsyg.2022.949881>.