

On the Definition of Prescriptive Annotation Guidelines for Language-Agnostic Subjectivity Detection

Federico Ruggeri^{1,*}, Francesco Antici¹, Andrea Galassi¹, Katerina Korre², Arianna Muti² and Alberto Barrón-Cedeño²

¹Department of Computer Science and Engineering (DISI), University of Bologna, Italy

²Department of Interpreting and Translation (DIT), University of Bologna, Italy

Abstract

Defining subjectivity indicators without relying on domain-specific assumptions or incurring interpretation biases is a well-known challenge. To account for these limitations, recent work is shifting toward annotation procedures for subjectivity detection that are not limited to language-specific cues. Nonetheless, developing a rigorous methodology to address edge cases and annotators' bias, while maintaining desired properties like language agnosticism, is yet an open problem. In this work, we rely on the prescriptive annotation paradigm and propose a methodology based on three key aspects. We present a case study on subjectivity detection for fact-checking in English and Italian news to evaluate the efficacy of the proposed methodology and discuss the open challenges.

Keywords

Subjectivity Detection, Annotation Guidelines, Natural Language Processing, Fact-Checking

1. Introduction

Subjectivity is a feature of language: when making an utterance, the speaker simultaneously expresses their position, attitude, and feelings towards the utterance, thus leaving their own mark [1]. Subjectivity Detection (SD) is the task of distinguishing objective content from subjective one. Previous SD approaches can be divided into syntactic and semantic [2]. The first category relies on keyword spotting [3, 4] or lexicons [5, 6, 7] as standard practice. However, these solutions are known to be language-specific unless some intermediate lossy translation procedure is considered [8]. Likewise, lexicon-based approaches require an external knowledge base which limits their applicability. In contrast, semantic approaches tackle SD via statistical [9, 10] or neural [11, 12, 13] methods for text representation by relying on labeled training corpora. This requirement is either addressed by considering domain-specific assumptions [9]

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M Litvak (eds.): *Proceedings of the Text2Story'23 Workshop, Dublin (Republic of Ireland), 2-April-2023*


*Corresponding author.

✉ federico.ruggeri6@unibo.it (F. Ruggeri); francesco.antici@unibo.it (F. Antici); a.galassi@unibo.it (A. Galassi); aikaterini.korre2@unibo.it (K. Korre); arianna.muti2@unibo.it (A. Muti); a.barron@unibo.it (A. Barrón-Cedeño)

🆔 0000-0002-1697-8586 (F. Ruggeri); 0000-0002-1125-0588 (F. Antici); 0000-0001-9711-7042 (A. Galassi); 0000-0002-9349-9554 (K. Korre); 0000-0002-3387-6557 (A. Muti); 0000-0003-4719-3420 (A. Barrón-Cedeño)



© 2023 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)

or designing annotation guidelines [11, 14, 15, 16].

Despite their independence from linguistic tools and allowing cross-lingual applicability with minor efforts [16, 17, 18], semantic approaches face a crucial yet demanding issue: the perception of subjectivity is itself subjective [2] and, thus, it is affected by interpretation bias [19], annotation ambiguity, and edge cases. As a result, defining practical, non-language-specific, and largely applicable annotation guidelines is a well-known challenge [15].

In this work, we adopt a prescriptive approach [20] and frame SD for a specific task to downplay annotation ambiguity [21], describing a method for the development of task-oriented annotation guidelines based on three key aspects: schematic case-based guidelines, iterative refinement, and reliable annotation. We also consider a preliminary case study on fact-checking to empirically evaluate the proposed methodology and elaborate on the encountered open challenges.

2. Methodology

We identify three key aspects for developing task-oriented SD annotation guidelines. We follow the prescriptive paradigm [20] to impose a specific and consistent conceptualization of subjectivity for annotation.

Schematic case-based guidelines. Given a task that partially relies on SD, it is necessary to define subjectivity according to the task’s objectives. It is, therefore, necessary to define annotation guidelines that are schematic and based on specific real cases. This formulation is less sensitive to domain- or language-specific cues and eases the annotators’ training process. Moreover, these properties could foster collecting large corpora for SD based on annotation guidelines rather than relying on domain-dependent assumptions [22].

Iterative refinement. Agreeing on a set of validated annotation guidelines is a collaborative refinement process. Such a process has the objective of discovering annotation edge cases, i.e. instances that are not covered by annotation guidelines resulting in high inter-annotator disagreement. Indeed, a preliminary version of annotation guidelines is unlikely to thoroughly cover all possible cases. For this reason, guideline refinement is an iterative process consisting of multiple annotation pilot studies since edge case discovery depends on the nature of sampled annotation data [23]. The pilot studies are designed to instruct annotators and reach a common set of validation annotation guidelines [24], and are iterated until a sufficient level of agreement is reached [25]. This formulation is in line with the prescriptive paradigm [20], where annotator disagreement is a call to action to refine annotation guidelines.

Reliable annotation. The last key aspect concerns the data annotation task. First, annotators are provided with refined annotation guidelines to instruct them. Second, text instances are assigned to multiple annotators to downplay the impact of noisy labels and annotators’ bias [19]. This process allows for discriminating edge cases from instances with a unanimous or almost perfect agreement. Tracking of individual annotations per instance is considered a measure of quality assurance [20, 26]. Eventually, labels can be aggregated via voting strategies for training

machine learning models [27]. In case of disagreement, a discussion phase among annotators takes place to agree on a solution. An additional annotator to label these instances is considered if an agreement is not reached. To address the problem of noisy labels, it is possible to discard those assigned by annotators that strongly disagree with each other [28] and explicitly report for which instances the discussion phase did not solve ambiguities [29, 30].

3. Discussion and Open Challenges

We elaborate on the presented methodology by discussing a case study on fact-checking. We consider a pipeline for fact-checking where SD is performed to discriminate between objective sentences that can be directly verified and subjective sentences that must be processed or rewritten to extract the objective claim or information. The detection and processing of subjective content have the final purpose of creating an objective narrative upon which fact-checking relies [31]. We consider the task of labeling sentences in English and Italian news articles targeting ongoing controversial topics, such as political affairs, Covid-19, civil rights, and economics (see Appendix A).

We initially design a set of preliminary annotation criteria suitable for fact-checking purposes (see Appendix B). These guidelines are mainly derived from existing work on SD on related domains [11, 32]. We recruit six human annotators with native or near-native knowledge of the English and Italian languages. After two annotation pilot studies, annotators agree on a common set of annotation criteria. We keep track of inter-annotator agreement (IAA) over pilot studies to validate their efficacy. In particular, the average Cohen’s kappa over annotator pairs is 0.39 (fair agreement) and 0.53 (moderate agreement) for the first and second pilot studies, respectively. We consider both Italian and English annotations when computing the IAA and observe comparable results between languages. The observed 14% gain between the two studies denotes a significant improvement in the annotation criteria.

During the pilot studies, we discuss the importance of contextual information (Section 3.1) for annotation and address several edge cases (Section 3.2). These observations are consistent in both languages, proving the efficacy of our methodology regardless of the language.

3.1. Annotating with Context

The lack of context may lead to ambiguous annotation cases, depending on the chosen input granularity [33, 34]. In our setting, we consider sentence-level granularity as common practice [31]. This choice represents a suitable testing ground for evaluating context importance given the limited scope of a sentence. For this purpose, in the second pilot study, we arrange annotators into two groups. Half of them label input sentences in order of appearance, while the remaining half labels sentences in random order, neglecting any contextual information as done in the first pilot study. We observe a 0.38 and 0.53 average Cohen’s kappa over annotator pairs for the context and non-context groups, respectively.

Our findings contrast the results of Ljubešić et al. [35], suggesting that context may be useful only in certain tasks or specific scenarios. Moreover, we identify two additional reasons in favor of a non-contextual annotation formulation. First, the use of context leads to an increased

Table 1

Example of edge cases encountered in our case study.

(a) Emotions	<i>He looked like he was on the verge of crying.</i>
(b) Quotes	<i>“Crosbie is an extremely violent man who has no place in society, and we welcome the jury’s verdict today.”</i>
(c) Intensifiers	<i>Recognising that, last Friday the US announced a further \$600m of military aid to Ukraine, including more Himars rockets that have so damaged Moscow’s logistics and its ability to resist.</i>
(d) Speculations	<i>Putin will hope to sow uncertainty in the eyes of policymakers’ meetings in New York.</i>

annotator’s workload. Consequently, it negatively affects the applicability of annotation guidelines to multiple scenarios. Second, contextual information may not be available in certain domains and settings, as in tweets [2]. These observations and the higher IAA suggest that a non-contextual annotation for SD is a preferred formulation.

3.2. Edge Cases

During our pilot studies, we identify four edge cases, as reported in Table 1.

Emotions. Statements carrying emotions convey a subjective point of view [36, 37] but they cannot be verified or confuted by a fact-checking system since they are based on the author’s beliefs and sensations only. Since it is impossible to provide such information in a more objective form, we label these statements as objective.

Quotes. In news sources, authors frequently use quotes to support their thesis. Even if the quoted content may be subjective, the task concerns detecting subjectivity only for the article’s author. For this reason, we label quoted content as objective.

Intensifiers. We identify intensifiers as indicators of subjectivity since their presence could be symptomatic of the author’s personal point of view. For example, in Table 1 (c) it is difficult to state if the expression “*so damaged*” conveys the author’s personal point of view or, rather, is descriptive and can be re-formulated as “*that have in this way damaged*”.

Speculations. Annotators often struggle to judge implicit statements without leveraging their own interpretation bias [38]. We consider speculation as a subjectivity indicator, since authors make use of it to allude to their own interpretation of events and consequences. The expression “*will hope to sow uncertainty*” in Table 1 (d) is an example.

4. Conclusions

We have presented our ongoing work on developing annotation guidelines for task-oriented SD. In particular, we introduced a methodology based on the prescriptive paradigm [20] to provide a task-specific definition of subjectivity via schematic and language-independent annotation criteria. These criteria are developed to cover annotation edge cases and downplay annotators’ interpretation biases. The application of our methodology to a preliminary case study on fact-checking in two different languages allowed us to reduce the ambiguity of the annotation by identifying edge cases and addressing them through the definition of specific guidelines. In future works, we will extend our approach to further languages.

Acknowledgments

This work has been partly funded by the European Union’s Horizon 2020 Research and Innovation programme under grant agreement 101017142 (“StairwAI: Stairway to AI”) and partly funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 (“FAIR - Future Artificial Intelligence Research” - Spoke 8 “Pervasive AI”), funded by the European Commission under the NextGeneration EU programme. A. Muti’s research is carried out under the project “DL4AMI–Deep Learning models for Automatic Misogyny Identification”, in the framework of *Progetti di formazione per la ricerca: Big Data per una regione europea più ecologica, digitale e resiliente*–Alma Mater Studiorum–Università di Bologna, Ref. 2021-15854. K. Korre’s research is carried out under the project “RACHS: Rilevazione e Analisi Computazionale dell’Hate Speech in rete”, in the framework of the PON programme FSE REACT-EU, Ref. DOT1303118.

References

- [1] L. Feng, On the subjectivity and intersubjectivity of language, in: *Communication and Linguistics Studies*, volume 6, 2020, pp. 1–5. doi:10.11648/j.cls.20200601.11.
- [2] I. Chaturvedi, E. Cambria, R. E. Welsch, F. Herrera, Distinguishing between facts and opinions for sentiment analysis: Survey and challenges, *Inf. Fusion* 44 (2018) 65–77. URL: <https://doi.org/10.1016/j.inffus.2017.12.006>. doi:10.1016/j.inffus.2017.12.006.
- [3] J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in: A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 486–497.
- [4] E. Riloff, J. Wiebe, Learning extraction patterns for subjective expressions, in: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 2003, pp. 105–112. URL: <https://aclanthology.org/W03-1014>.
- [5] N. Das, S. Sagnika, A subjectivity detection-based approach to sentiment analysis, in: D. Swain, P. K. Pattnaik, P. K. Gupta (Eds.), *Machine Learning and Information Processing*, Springer Singapore, Singapore, 2020, pp. 149–160.
- [6] H. Yu, V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, in: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP ’03*, Association for Computational Linguistics, USA, 2003, p. 129–136. URL: <https://doi.org/10.3115/1119355.1119372>. doi:10.3115/1119355.1119372.
- [7] J. Villena-Román, J. García-Morera, M. Á. G. Cumbreiras, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. U. López, Overview of TASS 2015, in: J. Villena-Román, J. García-Morera, M. Á. G. Cumbreiras, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. U. López (Eds.), *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015)*, Alicante, Spain, September 15, 2015, volume 1397 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 13–21. URL: <http://ceur-ws.org/Vol-1397/overview.pdf>.
- [8] F. Benamara, B. Chardon, Y. Mathieu, V. Popescu, Towards context-based subjectivity analysis, in: *Proceedings of 5th International Joint Conference on Natural Language*

- Processing, Asian Federation of Natural Language Processing, Chiang Mai, Thailand, 2011, pp. 1180–1188. URL: <https://aclanthology.org/I11-1132>.
- [9] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 2004, pp. 271–278. URL: <https://aclanthology.org/P04-1035>. doi:10.3115/1218955.1218990.
- [10] F. Sha, F. C. N. Pereira, Shallow parsing with conditional random fields, in: M. A. Hearst, M. Ostendorf (Eds.), Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada, May 27 - June 1, 2003, The Association for Computational Linguistics, 2003. URL: <https://aclanthology.org/N03-1028/>.
- [11] F. Antici, L. Bolognini, M. A. Inajetovic, B. Ivasiuk, A. Galassi, F. Ruggeri, Subjectivita: An italian corpus for subjectivity detection in newspapers, in: CLEF, volume 12880 of LNCS, Springer, 2021, pp. 40–52. URL: https://doi.org/10.1007/978-3-030-85251-1_4. doi:10.1007/978-3-030-85251-1_4.
- [12] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, The Association for Computer Linguistics, 2014, pp. 655–665. URL: <https://doi.org/10.3115/v1/p14-1062>. doi:10.3115/v1/p14-1062.
- [13] I. Chaturvedi, Y. Ong, I. Tsang, R. Welsch, E. Cambria, Learning word dependencies in text by means of a deep recurrent belief network, Knowledge-Based Systems 108 (2016). doi:10.1016/j.knosys.2016.07.019.
- [14] J. M. Wiebe, R. F. Bruce, T. P. O’Hara, Development and use of a gold-standard data set for subjectivity classifications, in: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, College Park, Maryland, USA, 1999, pp. 246–253. URL: <https://aclanthology.org/P99-1032>. doi:10.3115/1034678.1034721.
- [15] T. Wilson, J. Wiebe, Annotating opinions in the world press, in: Proceedings of the SIGDIAL 2003 Workshop, The 4th Annual Meeting of the Special Interest Group on Discourse and Dialogue, July 5-6, 2003, Sapporo, Japan, The Association for Computer Linguistics, 2003, pp. 13–22. URL: <https://aclanthology.org/W03-2102/>.
- [16] M. Abdul-Mageed, M. Diab, Subjectivity and sentiment annotation of Modern Standard Arabic newswire, in: Proceedings of the 5th Linguistic Annotation Workshop, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 110–118. URL: <https://aclanthology.org/W11-0413>.
- [17] I. Amini, S. Karimi, A. Shakery, Cross-lingual subjectivity detection for resource lean languages, in: Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Minneapolis, USA, 2019, pp. 81–90. URL: <https://aclanthology.org/W19-1310>. doi:10.18653/v1/W19-1310.
- [18] C. Banea, R. Mihalcea, J. Wiebe, Sense-level subjectivity in a multilingual setting, Computer Speech & Language 28 (2014) 7–19. URL: <https://www.sciencedirect.com/science/article/pii/S0885230813000181>. doi:<https://doi.org/10.1016/j.csl.2013.03.002>.

- [19] M. Geva, Y. Goldberg, J. Berant, Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 1161–1166. URL: <https://doi.org/10.18653/v1/D19-1107>. doi:10.18653/v1/D19-1107.
- [20] P. Röttger, B. Vidgen, D. Hovy, J. B. Pierrehumbert, Two contrasting data annotation paradigms for subjective NLP tasks, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 175–190. URL: <https://doi.org/10.18653/v1/2022.naacl-main.13>. doi:10.18653/v1/2022.naacl-main.13.
- [21] T. A. Wilson, Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states, University of Pittsburgh, 2008.
- [22] R. Satapathy, S. Pardeshi, E. Cambria, Polarity and subjectivity detection with multitask learning and BERT embedding, *Future Internet* 14 (2022) 191. URL: <https://doi.org/10.3390/fi14070191>. doi:10.3390/fi14070191.
- [23] V. K. Pradhan, M. Schaekermann, M. Lease, In search of ambiguity: A three-stage workflow design to clarify annotation guidelines for crowd workers, *Frontiers Artif. Intell.* 5 (2022) 828187. URL: <https://doi.org/10.3389/frai.2022.828187>. doi:10.3389/frai.2022.828187.
- [24] R. Artstein, Inter-annotator agreement, *Handbook of linguistic annotation* (2017) 297–313.
- [25] E. Musi, D. Ghosh, S. Muresan, Towards feasible guidelines for the annotation of argument schemes, in: Proceedings of the Third Workshop on Argument Mining (ArgMining2016), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 82–93. URL: <https://aclanthology.org/W16-2810>. doi:10.18653/v1/W16-2810.
- [26] M. Teruel, C. Cardellino, F. Cardellino, L. A. Alemany, S. Villata, Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA), 2018. URL: <http://www.lrec-conf.org/proceedings/lrec2018/summaries/1048.html>.
- [27] A. T. Nguyen, B. Wallace, J. J. Li, A. Nenkova, M. Lease, Aggregating and predicting sequence labels from crowd annotations, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 299–309. URL: <https://aclanthology.org/P17-1028>. doi:10.18653/v1/P17-1028.
- [28] J. Amidei, P. Piwek, A. Willis, Identifying annotator bias: A new IRT-based method for bias identification, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 4787–4797. URL: <https://aclanthology.org/2020.coling-main.421>. doi:10.18653/v1/2020.coling-main.421.

- [29] V. Basile, F. Cabitza, A. Campagner, M. Fell, Toward a perspectivist turn in ground truthing for predictive computing, *CoRR abs/2109.04270* (2021).
- [30] G. Abercrombie, V. Basile, S. Tonelli, V. Rieser, A. Uma (Eds.), *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, European Language Resources Association, Marseille, France, 2022. URL: <https://aclanthology.org/2022.nlperspectives-1.0>.
- [31] Z. Guo, M. Schlichtkrull, A. Vlachos, A Survey on Automated Fact-Checking, *Transactions of the Association for Computational Linguistics* 10 (2022) 178–206. URL: https://doi.org/10.1162/tacl_a_00454. doi:10.1162/tacl_a_00454. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00454/1987018/tacl_a_00454.pdf.
- [32] L. de Saussure, P. Schulz, Subjectivity out of irony, *Semiotica* 2009 (2009) 397–416. URL: <https://doi.org/10.1515/SEMI.2009.018>. doi:doi:10.1515/SEMI.2009.018.
- [33] J. Pavlopoulos, J. Sorensen, L. Dixon, N. Thain, I. Androutsopoulos, Toxicity detection: Does context really matter?, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 4296–4305. URL: <https://aclanthology.org/2020.acl-main.396>. doi:10.18653/v1/2020.acl-main.396.
- [34] S. Menini, A. P. Aprosio, S. Tonelli, Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection, *CoRR abs/2103.14916* (2021). URL: <https://arxiv.org/abs/2103.14916>. arXiv:2103.14916.
- [35] N. Ljubešić, I. Mozetič, P. K. Novak, Quantifying the impact of context on the quality of manual hate speech annotation, *Natural Language Engineering* (2022) 1–14.
- [36] R. Mihalcea, C. Banea, J. Wiebe, Multilingual subjectivity and sentiment analysis, in: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Association for Computational Linguistics, Jeju Island, Korea, 2012, p. 4. URL: <https://aclanthology.org/P12-4004>.
- [37] K. Veronika, Subjectivity and emotions as sources of insight in an ethnographic case study: A tale of the field, *M@n@gement* 9 (2006) 117–135. URL: <https://management-aims.com/index.php/mgmt/article/view/4089>.
- [38] T. Caselli, V. Basile, J. Mitrović, I. Kartoziya, M. Granitzer, I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language, in: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 6193–6202. URL: <https://aclanthology.org/2020.lrec-1.760>.

Appendix

A. News Sources Considered

For our pilot studies, we consider the news sources reported in Table 2. For each study, we randomly sample up to six articles (~ 150 sentences on average). All the annotators label the sampled articles at the sentence level.

Table 2

Sources considered for the pilot studies.

English		Italian	
frontpagemag.com	shtfplan.com	fascinazione.it	ilfoglio.it
telegraph.co.uk	theguardian.com	avantionline.it	liberoquotidiano.it
vdare.com		avvenire.it	

B. Initial Draft of Annotation Guidelines

The initial set of annotation criteria for subjectivity detection states that a sentence is subjective if:

- (i) *it explicitly reports the personal opinion of its author;*
- (ii) *it contains sarcastic or ironic expressions;*
- (iii) *it contains exhortations or personal auspices;*
- (iv) *it contains discriminating or downgrading expressions;*
- (v) *it contains rhetorical figures explicitly made by its author to convey their opinion;*
- (vi) *it contains a conclusion made by its author that is drawn despite insufficient factual information.*

After the first pilot study, annotators identify and discuss two major edge cases: emotions and quotes. In particular, the following annotation criteria are added:

- (vii) *a sentence is objective when it describes the personal feelings, emotions or moods of its author, without conveying opinions on other matters;*
- (viii) *a sentence is objective if it expresses an opinion, claim, emotion, or a point of view that is explicitly attributable to a third-party (e.g., a person mentioned in the text). The presence of quotation marks (“ ”), when used to quote a third person (be it at the beginning of the sentence, at the end, or both), represents an explicit third-party opinion, even if it is not clearly stated in the sentence.*

Additionally, annotation criteria (i) is modified to explicitly address rhetorical questions: *rhetorical questions are considered as an expression of opinion.*

After the second pilot study, annotators identify and discuss two additional edge cases: speculations and intensifiers. In particular, the following annotation criteria are added:

- (ix) *a sentence is subjective if it contains intensifiers that can be attributed to its author to express their opinion.*

Moreover, annotation criteria (i) is modified to address speculations: *speculations that draw conclusions are considered opinions.*