Borrowing Words: Transfer Learning for Reported Speech Detection in Slovenian News Texts

Zoran Fijavž
Jožef Stefan Postgraduate International School
Peace Institute
Slovenia, Ljubljana
zoran.fijavz@mirovni-institut.si

Abstract

This paper describes the development of a reported speech classifier for Slovenian news texts using transfer learning. Due to a lack of Slovenian training data, multilingual models were trained on English and German reported speech datasets, reaching an F-score of 66.8 on a small manually annotated Slovenian news dataset and a manual error analysis was performed. While the developed model captures many aspects of reported speech, further refinement and annotated data would be needed to reliably predict less frequent instances, such as indirect speech and nominalizations.

Keywords

reported speech, natural language processing, transfer learning, news analysis

1 Introduction

Reported speech, ubiquitous in literary and news texts, has clear lexical and syntactic patterns which may be reliably modeled via natural language processing (NLP) and may be useful for downstream tasks by drawing a distinction between source and background information. The paper applies transfer learning to extend reported speech classification to Slovenian news texts and provides a provisional classification model. A manual error analysis reveals the model's strengths and weaknesses, highlighting possible steps for further improvements.

2 Related Work

2.1 Role of Reported Speech

Reported speech is common in news texts, generally expressed as direct or indirect speech, with the former repeating the original utterance verbatim and the latter embedding it in a that-clause [18] (e.g., *Jimmy said: "Another systematic review would be great!"* and *Jimmy said that another systematic review would be great.*). More complex forms include mixed speech (*City officials rebuffed the accusations as "groundless and blatantly false".*) and reportative nominalizations with an analogous function as reported speech (*The speaker particularly emphasized the pressures on the media*) [7]. Around 50% of sentences in newspaper corpora may be attributed to a source in the text, predominantly through direct and indirect speech [17]. Verbs cue 96% of reported speech, followed by prepositional phrases (3%) [13]. Reported speech lends objectivity to statements [9], summarizes source statements [16], and is used in discourse analysis and communication studies to explore speaker representation by gender [1], institutional affiliations [8], and topic stances [15], or to distinguish between journalists' and sources' voices [11].

2.2 Existing Datasets and Modelling Approaches

Datasets with reported speech annotations mostly cotain literary or news texts. Key corpora include RiQuA [12], SLäNDa 2.0 [19], Redewiedergabe [3], QUAC [14], PolNeAR [10], Quotebank [21], and STOP [22]. RiQuA and Redewiedergabe are the largest annotated corpora, covering English and German 19th century texts. QUAC contains 212 annotated articles from the Portuguese newspaper Público, while Quotebank spans 162 million news articles with automatic annotations. PolNeAR, consisting of 1,028 news articles, includes attribution annotations, which include and exceed the definition of reported speech. A summary of the datasets is provided in Table 1.

The corpora differ in annotation complexity and size. They are mostly monolingual, warranting the used cross-lingual transfer learning for low-resource languages by employing multilingual models such as mBERT [6] and XLM-R [4]. Narrower multilingual models, such as CroSloEngual BERT, often outperform broader ones [20]. Reported speech modeling may be operationalized as speaker or quotation detection tasks [23, 17]. Simplifying the task to sentence-level classification is warranted by the fact news (unlike literary texts) rarely mix statements by sources and authors in the same sentence and can improve classification reliability at the expense of detailed aspects of reported speech [17] and simplify the annotation structure. Missing fine-grained outputs, such as speakers and boundaries of reported and reporting clauses, may thus be an acceptable trade-off for NLP-based content analysis in news texts. A systematic review of such approaches points to the limits resulting from a low number of features with no guarantee of reliable (joint) prediction, which preclude drawing rich conclusions expected from the method's manual counterpart [2].

3 Experimental Setting

3.1 Task Overview

We treated reported speech as a sentence-level classification task. Sentence splitters were applied to existing datasets, and binary labels were assigned by matching annotated spans with the split sentences. Reported speech sub-types were unified under a single label, joining the annotation schemes of individual datasets. A Slovenian dataset of 10 news texts was manually annotated at the sentence level. The datasets were split into training, evaluation, and test sets to train multilingual pretrained models. For CroSloEngual BERT, preprocessing also involved machine translating the German training data into English. The model outputs were binary labels indicating reported speech, used to calculate F-scores on the test data. A manual error analysis was performed on

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.21

Table 1: Summary of Datasets' Characteristics.

Corpus	Type	Annotations	Language	Sentence No.	Role	Positive Class
RiQua	fiction	direct and indirect speech, cues, speakers, addressees	English	38,610	72% train, 18% development, 10% test	48%
Redewiedergabe	fiction, news	direct, indirect, free indirect and reported speech, speaker, cues	German	24,033	76% train, 16% development, 9% test	33%
Quotebank (manual)	news	speaker, direct speech	English	9,071	test	30%
QUAC	news	speaker, direct speech	Portuguese	11,007	test	11%
PolNeAR	news	speaker, cues, attributions	English	34,153	test	59%
Slovenian parliamentary news	news	sentence-level binary labels	Slovenian	744	test	43%

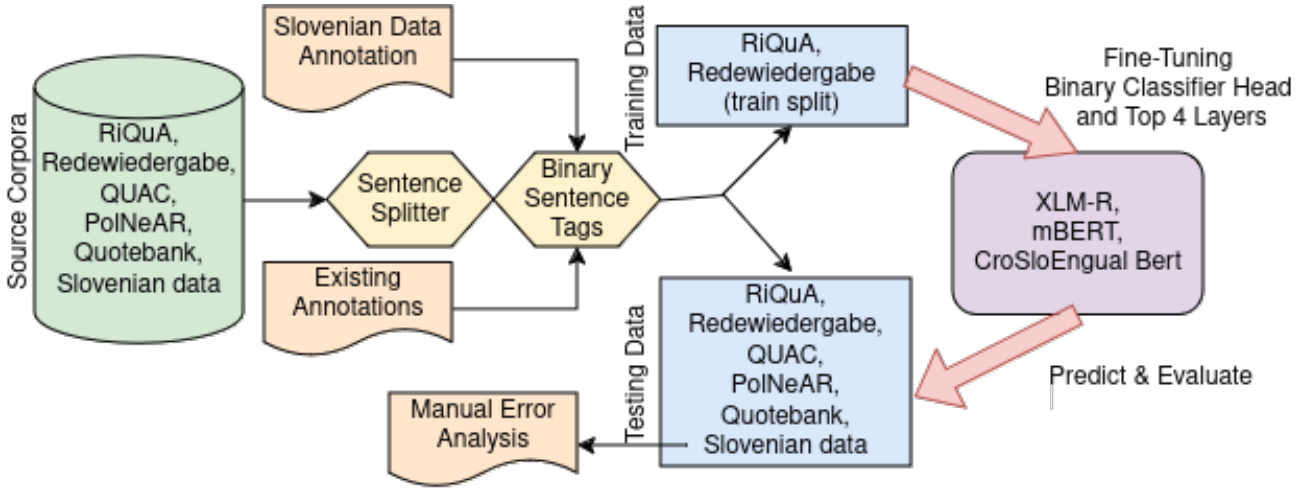

Figure 1: Flowchart of Data Preprocessing, Model Training and Evaluation Processes for Sentence-Level Reported Speech Classification.

the best model's outputs for Slovenian. Preprocessing, training, and evaluation steps are visualized in 1.

3.2 Training and Test Data

Our experiments were based on existing annotated reported speech datasets and a small Slovenian dataset. The training data included sections from RiQuA and Redewiedergabe, both large datasets with labels for direct and indirect speech. For CroSloEngual BERT training, the Redewiedergabe data was machine translated into English. Testing was conducted on the test sections of RiQuA, Redewiedergabe, the entire Portuguese corpus QUAC, and the manually annotated portion of the English Quotebank corpus. Additionally, we manually annotated 10 Slovenian news articles from RTV Slovenia. The datasets are summarized in Table 1.

The Slovenian dataset comprised 10 parliamentary news texts, covering various reporting strategies. Retrieved articles were split into sentences and annotated. Sentences were considered reported speech if they included direct or indirect speech cued by a reporting clause or prepositional phrase. We excluded nominalizations and phrasal quotes (e.g., *They emphasized the pressures on the media and the "illegal non-funding of the Press Agency."*) as well as implied quotes (e.g., *There will be more than 300,000 recipients, he emphasized. 169 million euros will have to be paid out.*).

3.3 Evaluation Procedure

The models' performance on the test datasets was calculated with an F-score. A baseline of assigning a positive label to all examples was calculated for all test datasets. The models' results on the test datasets were compared with a Friedman's test as suggested in the literature [5].

The best Slovenian model's predictions were reviewed with close reading. The error typology consisted of direct speech, indirect speech, speech fragments, annotation errors, annotation errors and *unrelated* and *other* tags. *Direct speech fragments* were sentences part of multi-sentence direct speech quotations. *Annotation errors* were examples with annotations inconsistent with the definition described in *Section 3.2.* For *unrelated* examples, close reading revealed no clear misclassification cause. *Other* was used for examples that did not fit any of the mentioned categories.

3.4 Training Settings

XLM-R and mBERT were used as base models with the default training settings from the *transformers* library with the exception of using 16 gradient accumulation steps and freezing the bottom 8 layers of all models. The latter reduces the training time without significant performance drops (Kovaleva idr., 2019; Merchant idr., 2020). Additionally, a Slovenian-Croatian-English BERT model was trained on English machine-translated data from Redewiedergabe.

4 Results

4.1 Model Results

The model performance varies based on the congruence between the language and precise task definitions in each dataset. The differences between model predictions were not statistically significant (χ^2_F = 9.66; df = 5; n = 8; p = 0.14) so post-hoc tests were not performed. As Table 2 demonstrates, the XLM-R model trained on both RiQuA and Redewiedergabe performed well across the datasets with an F-score of 80.5 and 77.6 on the Redewiedergabe and RiQuA test set, respectively. The high results from training on combined data suggests the RiQuA and Redewiedergabe datasets may benefit from additional or complementary data, at least when using cross-lingual transfer learning. The most successful strategy for Slovenian data was training on RiQuA and English machine-translated Redewiedergabe data using the CroSloEngual BERT model, reaching a F-score of 66.8. We did not evaluate the impact of using translated training data with mBERT and XLM-R.

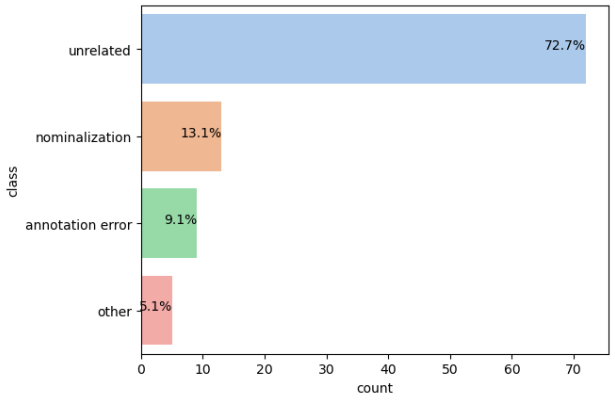

Figure 2: False Positives from the CroSloEngual BERT Classifier.

4.2 Error Analysis Results

The results from CroSloEngual BERT on Slovenian data were analyzed further. False positives were more common than false negatives, representing 23.4% and 9.8% of all examples (n = 744), respectively. Close reading of a sample of 100 false positives did not show a definite pattern for most (72.9%) of them. These examples were clearly unrelated to reported speech, although some did include words lexically related to reporting verbs (e.g. *The proposed law is still under discussion*). The second category of false positives were nominalizations of reported statements (13.1%) not included in our annotation schema. The final source of false positives were annotation errors consisting of wrongly

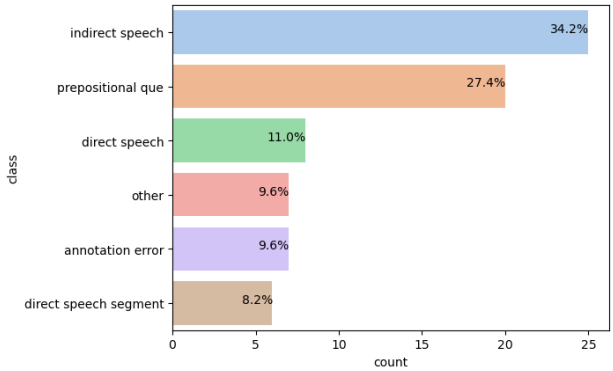

Figure 3: False Negatives from the CroSloEngual BERT Classifier.

unmarked examples of direct or indirect speech (9.1%). The distribution of categories identified in the sample of false positives are illustrated in Figure 2. The most common errors in the 73 false negative examples were instances of indirect speech (34.2% of false negatives) and prepositional queing of reported speech (27.4%). The remainder were instances of direct speech, direct speech fragments and annotation errors representing 11%, 8.2% and 9.6% of the false negatives, respectively. The annotation errors included nominalizations and statements reported as adjective complements (*The speaker was happy that the provisions were accepted)* not included in our annotation schema. Figure 3 summarizes the identified false negative categories .

5 Discussion

This paper presents the development of a reported speech classifier, tested through a small annotated Slovenian dataset and manual error analysis. Cross-lingual transfer learning from the annotated RiQuA and Redewiedergabe datasets achieved an F-score of 66.8 on a small manually annotated dataset of Slovenian news of parliamentary sessions using the base CroSloEngual model with RiQuA and English machine-translated Redewiedergabe training data [1]. This these results corroborate the observation that language models trained on a limited number of languages may outperform less specialized ones such as mBERT and XLM-R [20]. The major source of errors were false positives (23.4% of all sentences) for which no systematic pattern was discernible in the majority (72.9%) of examples. Instances of indirect speech and prepositional queing of statements were overrepresented in the false negatives, accounting for 61.6% of false negatives. Although rare, nominalizations were present in both false positives and false negatives and should be considered in future annotation guidelines. These obeservations indicate reported speech classifiers may benefit form approaches for addressing imbalanced classes.

6 Conclusion

This study developed a sentence-level reported speech classifier for Slovenian news texts using cross-lingual transfer learning. By leveraging existing multilingual models (mBERT, XLM-R, and CroSloEngual BERT) with the English and German datasets RiQuA and Redewiedergabe, we demonstrated that sentence-level

[1]The fine-tuned CSE model is available on the Hugging Face Hub under the name *zo-fi/rep-sp-CSE-rwg-riq*.

Table 2: Model Performances across Datasets (F-scores).

	Redewiedergabe	RiQuA	PolNeAR	QUAC	Quotebank	Slovenian dataset
Positive by default	52.1	60.6	74.2	19.5	45.8	60.3
mBERT+Both	77.5	77.4	73.1	40.5	53.5	63.2
mBERT+RiQuA	68.2	76.9	72.6	31.1	52.6	39.1
mBERT+RWG	78.4	70.4	65.5	43.4	49.1	63.2
XLM-R+Both	80.5	77.6	70	38.8	57.7	63.2
XLM-R+RiQuA	66.6	76.7	73.6	25.5	53.7	60.3
XLM-R+RWG	80.9	70.7	66.4	43.9	50	63.2
CroSloEngBERT+Both+MT	54	76.6	73	24	52.5	66.8

classification can detect some aspects of reported speech in Slovenian. However, the performance estimates are limited due to the small size of the Slovenian testing set and the limited definition used for the annotations. Future research should focus on developing a Slovenian annotated dataset, refining the annotation schema for multiple use cases, and exploring additional modeling features such as encoding broader sentence contexts. This work contributes a provisional tool for computational discourse analysis of Slovenian media texts. Further development is necessary for its application in more nuanced tasks.

Acknowledgements

This work was supported by the Slovenian Research Agency grants via the core research programs Equality and Human Rights in the Times of Global Governance (P5-0413) and Hate Speech in Contemporary Conceptualizations of Nationalism, Racism, Gender and Migration (J5-3102).

References

[1] Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021. The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media. *PLOS ONE*, 16, 1, (Jan. 29, 2021), e0245533. DOI: 10.1371/journal.pone.0245533.

[2] Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16, 1, (Jan. 2, 2022), 1–18. DOI: 10.1080/19312458.2021.2015574.

[3] Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Nicoletta Calzolari et al., editors. European Language Resources Association, 803–812. ISBN: 979-10-95546-34-4. https://aclanthology.org/2020.lrec-1.100.

[4] Alexis Conneau et al. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors. Association for Computational Linguistics, 8440–8451. DOI: 10.18653/v1/2020.acl-main.747.

[5] Janez Demšar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *The Journal of Machine Learning Research*, 7, (Dec. 1, 2006), 1–30.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. NAACL-HLT 2019. Jill Burstein, Christy Doran, and Thamar Solorio, editors. Association for Computational Linguistics, 4171–4186. DOI: 10.18653/v1/N19-1423.

[7] Gabriel Dvoskin. 2020. Reported speech and ideological positions: the social distribution of knowledge and power in media discourse. *Bakhtiniana: Revista de Estudos do Discurso*, 15, 193–213.

[8] Zoran Fijavž and Darja Fišer. 2021. Citatnost in reprezentacija v spletnem migracijskem diskurzu. In *Sociolingvistično iskrenje*. Maja Bitenc, Marko Stabej, and Žejn Andrejka, editors. Založba Univerze v Ljubljani. Retrieved Apr. 3, 2024 from https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/259/370/6011.

[9] Elizabeth Holt. 1996. Reporting on Talk: The Use of Direct Reported Speech in Conversation. *Research on Language and Social Interaction*, 29, 3, (July 1, 1996), 219–245. DOI: 10.1207/s15327973rlsi2903_2.

[10] Edward Newell, Drew Margolin, and Derek Ruths. 2018. An Attribution Relations Corpus for Political News. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. Nicoletta Calzolari et al., editors. European Language Resources Association (ELRA). Retrieved Apr. 10, 2024 from https://aclanthology.org/L18-1524.

[11] Mojca Pajnik and Marko Ribać. 2021. Medijski populizem in afektivno novinarstvo: časopisni komentar o »begunski krizi«. *Javnost - The Public*, (Dec. 14, 2021). Retrieved Apr. 24, 2024 from https://www.tandfonline.com/doi/abs/10.1080/13183222.2021.2012943.

[12] Sean Papay and Sebastian Padó. 2020. RiQuA: A Corpus of Rich Quotation Annotation for English Literary Text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. LREC 2020. Nicoletta Calzolari et al., editors. European Language Resources Association, 835–841. ISBN: 979-10-95546-34-4. Retrieved Apr. 21, 2024 from https://aclanthology.org/2020.lrec-1.104.

[13] Silvia Pareti, Tim O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically Detecting and Attributing Indirect Quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2013. David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors. Association for Computational Linguistics, 989–999. Retrieved Apr. 17, 2024 from https://aclanthology.org/D13-1101.

[14] Marta Ercília Mota Pereira Quintão. 2014. Quotation A ttribution for Portuguese News Corpora. In Retrieved Apr. 21, 2024 from https://www.semanticscholar.org/paper/Quotation-A-ttribution-for-Portuguese-News-Corpora-Quint%C3%A3o/69fea7d030d5e71b973ec67aa897a7c9aadadac2.

[15] Masaki Shibata. 2023. Dialogic Positioning on Pro-Whaling Stance: A Case Study of Reported Speech in Japanese Whaling News. *Japanese Studies*, 43, 1, (Jan. 2, 2023), 71–90. DOI: 10.1080/10371397.2023.2191839.

[16] Michael Short. 1988. Speech presentation, the novel and the press. In *The Taming of the Text*. Willie Van Peer, editor. Routledge. ISBN: 978-1-315-54452-6.

[17] Alexander Spangher, Nanyun Peng, Jonathan May, and Emilio Ferrara. 2023. Identifying Informational Sources in News Articles. Version 1. DOI: 10.48550/ARXIV.2305.14904.

[18] Stef Spronck and Daniela Casartelli. 2021. In a manner of speaking: how reported speech may have shaped grammar. *Frontiers in Communication*, 6, 624486.

[19] Sara Stymne and Carin Östman. 2022. SLäNDa version 2.0: Improved and Extended Annotation of Narrative and Dialogue in Swedish Literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. LREC 2022. Nicoletta Calzolari et al., editors. European Language Resources Association, 5324–5333. Retrieved Apr. 21, 2024 from https://aclanthology.org/2022.lrec-1.570.

[20] Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Text, Speech, and Dialogue* (Lecture Notes in Computer Science). Petr Sojka, Ivan Kopeček, Karel Pala, and Aleš Horák, editors. Springer International Publishing, Cham, 104–111. ISBN: 978-3-030-58323-1. DOI: 10.1007/978-3-030-58323-1_11.

[21] Timoté Vaucher, Andreas Spitz, Michele Catasta, and Robert West. 2021. Quotebank: A Corpus of Quotations from a Decade of News. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. WSDM '21: The Fourteenth ACM International Conference on Web Search and Data Mining. ACM, 328–336. ISBN: 978-1-4503-8297-7. DOI: 10.1145/3437963.3441760.

[22] M. Wynne. 1996. Speech, Thought and Writing Presentation Corpus. Retrieved Apr. 21, 2024 from https://ora.ox.ac.uk/objects/uuid:6caa73c1-d283-4d51-a78f-55df69bae986.

[23] Dian Yu, Ben Zhou, and Dong Yu. 2022. End-to-End Chinese Speaker Identification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2022. Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors. Association for Computational Linguistics, 2274–2285. DOI: 10.18653/v1/2022.naacl-main.165.