Title: Matching claims appendix for "'Fact checking' fact checkers: A data-driven approach"
Authors: Sian Lee (1), Aiping Xiong (1), Haeseung Seo (1), Dongwon Lee (1)
Date: October 26th, 2023
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

Appendix C: Matching claims

Lim (2018) compared statements of PolitiFact and Washington Post Fact Checker (WPFC) and their ratings. The author manually collected 1,178 and 325 fact-checking articles about the 2016 U.S. presidential candidates' statements from PolitiFact and WPFC, respectively from September 2013 to November 2016. Then, two raters manually labeled whether the statements overlapped (i.e., same), were murky (i.e., similar), or neither based on the title of the article. They found that there were 77 overlapping (i.e., matching) claims.

The labeled data was utilized to identify the optimal model for the automatic matching of claims. Our primary focus was on identifying overlapping cases. Thus, we dropped the murky label and re-labeled them as a binary class of either overlapping or non-overlapping. Given that PolitiFact possessed a larger dataset than WPFC, we employed the PolitiFact dataset as our baseline for comparison. Then, we tried Count Vectorizer (i.e., bag-of-words, Qader et al., 2019), term frequency-inverse document frequency (TF-IDF) Vectorizer (Kaur et al., 2020), and sentence BERT for sentence (Reimers & Gurevych, 2019), and varied the thresholds, x, in 0.05 intervals ranging from 0 to 1 to identify the optimal approach for determining matching claims. For the distance metric, we used cosine similarity. For the performance metric, we used F1-score for the positive class (i.e., overlapping) because we had fewer positive cases (*N* = 77) than negative cases (*N* = 1,101).

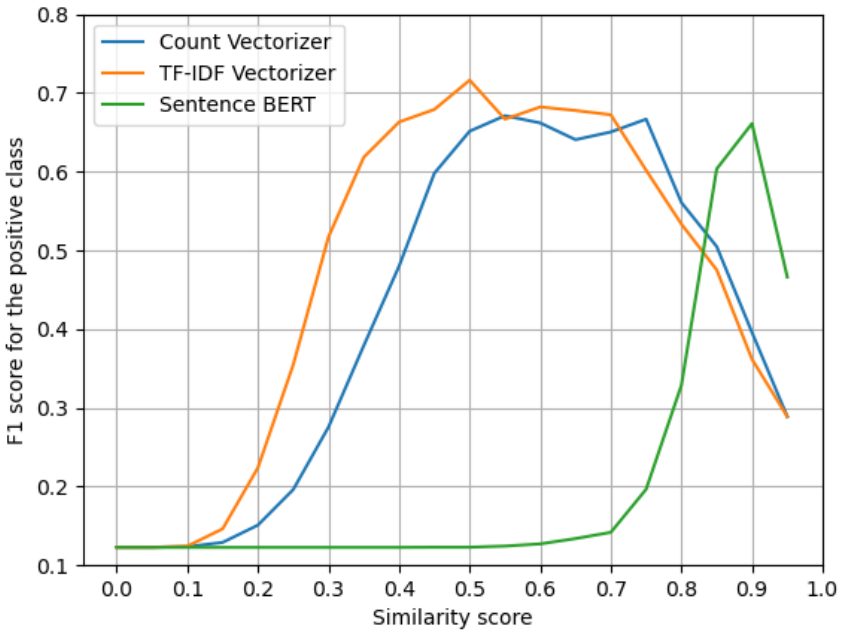

Figure A2. F1 score for the positive class. *F1 score for overlapping cases with Count Vectorizer, TF-IDF vectorizer, and sentence BERT. We tested different cosine similarity scores ranging from 0 to 1 with 0.05 intervals. TF-IDF with a cosine similarity score of 0.5 gives the best performance.*

Figure A2 shows the F1 scores for the different word-embedding methods and found that the TF-IDF method achieved the best performance, with a threshold of 0.5, based on the labeled data. Therefore, to

identify matching claims between different fact checkers, we used a TF-IDF and cosine similarity threshold of 0.5 and applied it to each claim. If any of the two claims showed the cosine similarity x $\geq$ 0.5, then we labeled them as *matching claims*. If there were multiple claims showing cosine similarity x $\geq$0.5, then we selected the claim which gave the highest similarity to the matching claim. Specifically, for each claim "A" fact-checked by Snopes, we identified a matching claim "B" by PolitiFact with the highest similarity score above 0.5. To ensure a comprehensive analysis, we also conducted the same analysis in reverse order, starting from each claim published by PolitiFact and comparing it with claims by Snopes. Table A1 shows the results of claims matching, and Figure A3 displays a confusion matrix that illustrates the performance of our model. Overall, the results showed an accuracy of 0.96. Moreover, the model achieved an F1-score of 0.72 for the *claims matched* label.

Table A1. *Automated claims matching results.*

Model	Accuracy	Precision	Recall	F1-score	Support
Claims not matched		0.98	0.98	0.98	1101
Claims matched		0.75	0.69	0.72	77
Overall	0.96				1178
Macro avg.		0.86	0.84	0.85	1178
Weighted avg.		0.96	0.96	0.96	1178

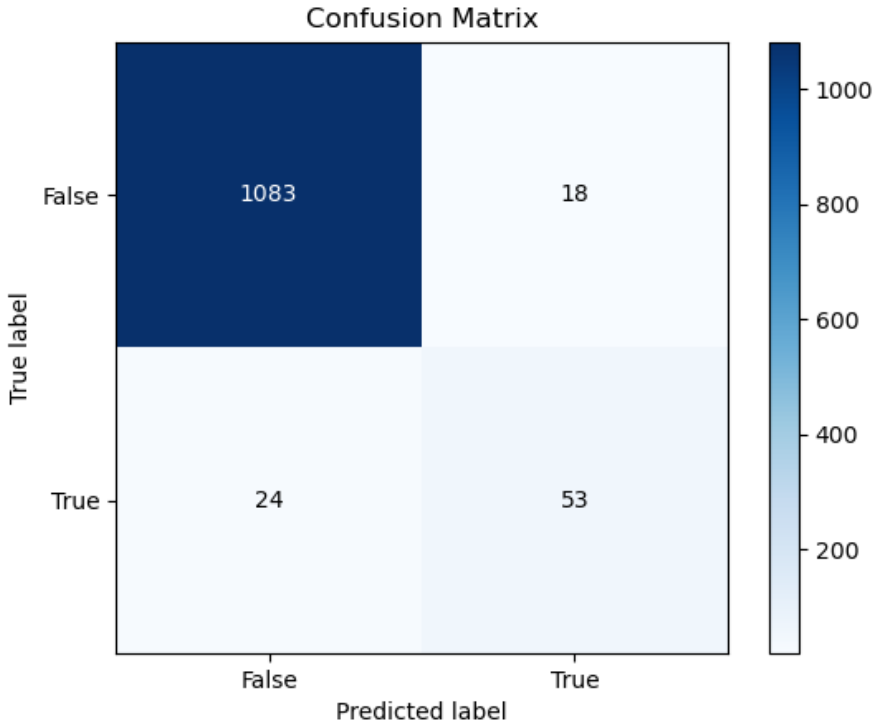

Figure A3. Confusion matrix. *With TF-IDF and cosine similarity score of 0.5.*