# Automatic Fake News Detection: Are current models “fact-checking” or “gut-checking”?

**Ian Kelk** **Benjamin Basseri** **Wee Yi Lee** **Richard Qiu** **Chris Tanner**

Harvard University
{iak415@g,basseri@cs50,wel390@g}.harvard.edu
{rqiu@college,christanner@g}.harvard.edu

## Abstract

Automatic fake news detection models are ostensibly based on logic, where the truth of a claim made in a headline can be determined by supporting or refuting evidence found in a resulting web query. These models are believed to be reasoning in some way; however, it has been shown that these same results, or better, can be achieved without considering the claim at all – only the evidence. This implies that other signals are contained within the examined evidence, and could be based on manipulable factors such as emotion, sentiment, or part-of-speech (POS) frequencies, which are vulnerable to adversarial inputs. We neutralize some of these signals through multiple forms of both neural and non-neural pre-processing and style transfer, and find that this flattening of extraneous indicators can induce the models to actually require both claims and evidence to perform well. We conclude with the construction of a model using emotion vectors built off a lexicon and passed through an “emotional attention” mechanism to appropriately weight certain emotions. We provide quantifiable results that prove our hypothesis that manipulable features are being used for fact-checking.

## 1 Introduction

Recent events such as the last two U.S. presidential elections have been greatly affected by fake news, defined as “fabricated information that disseminates deceptive content, or grossly distort actual news reports, shared on social media platforms” (Allcott and Gentzkow, 2017). In fact, the World Economic Forum 2013 report designates massive digital misinformation as a major technological and geopolitical risk (Bovet and Makse, 2019). As daily social media usage increases (Statista Research Department, 2021), manual fact-checking cannot keep up with this deluge of information.

Automatic fact-checking models are therefore a necessity, and most of them function using a system of *claims* and *evidence* (Hassan et al., 2017). Given a specific claim, the models use external knowledge as evidence. Typically, a web search query is treated as the claim, and a subset of the top search results is treated as the evidence. There is an implicit assumption that the fact-checking models are reasoning in some way, using the evidence to confirm or refute the claim. Recent research (Hansen et al., 2021) found this conclusion may be premature; current models can show improved performance when considering evidence alone, essentially fact-checking an unasked question. While this might seem reasonable given that the evidence is conditioned on the claims by the search engine, this can be exploited as illustrated in Figure 1, which shows that evidence returned using a ridiculous claim can still appear reasonable if we view the evidence alone without the claim. Furthermore, textual entailment requires both a text and a hypothesis; if we have a result without a hypothesis, we are performing a different, unknown task.

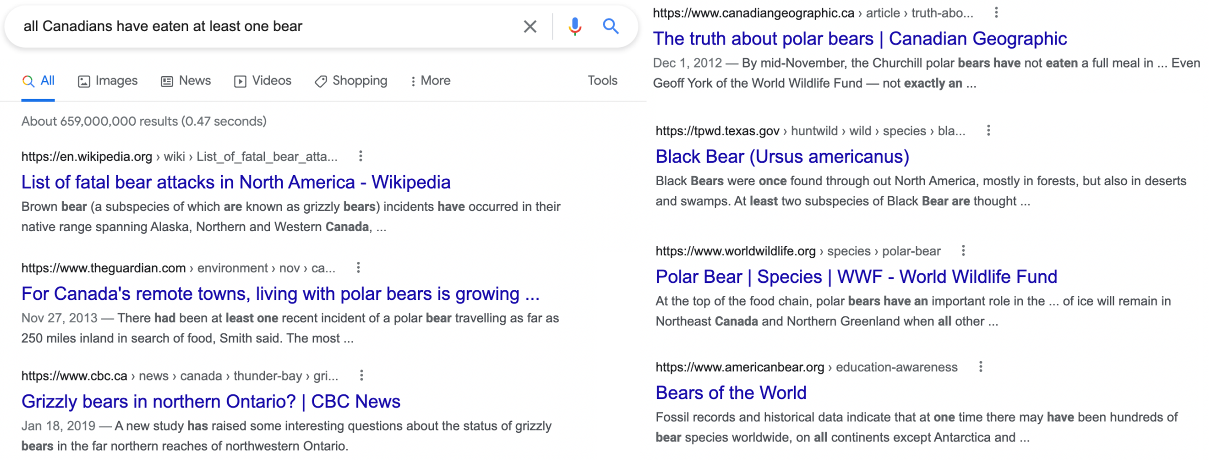
This finding indicates a problem with current automatic fake news detection, signaling that the models rely on features in the evidence typical to fake news, rather than using entailment. Since most automated fact-checking research is primarily concerned with the accuracy of the results, rather than addressing *how* the results are achieved, we propose a novel investigation into these models and their evidence. We use a variety of pre-processing steps, including neural and non-neural ones, to attempt to reduce the affectations common in evidence:

- Stemming, stopword removal, negation, and POS-filtering (Babanejad et al., 2020).
- Style transfer neural models using the *Styleformer* model to perform **informal-to-formal** and **formal-to-informal** paraphrasing methods (Li et al., 2018; Schmidt, 2020).

We also develop our own BERT-based model as an extension of the *EmoCred* system (Giachanou

*Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 29 - 36
May 26, 2022 ©2022 Association for Computational Linguistics

Figure 1: An example of why evidence alone does not suffice in identifying fake news, despite the evidence being conditioned on the claim as a search-engine query. Although the returned evidence appearing reputable, it is clear that it has little relevance to deciding the veracity of the claim that "all Canadians have eaten at least one bear."



all Canadians have eaten at least one bear

About 659,000,000 results (0.47 seconds)

https://en.wikipedia.org › wiki › List_of_fatal_bear_atta...
List of fatal bear attacks in North America - Wikipedia
Brown **bear** (a subspecies of which **are** known as grizzly **bears**) incidents **have** occurred in their native range spanning Alaska, Northern and Western **Canada**, ...

https://www.theguardian.com › environment › nov › ca...
For Canada's remote towns, living with polar bears is growing ...
Nov 27, 2013 — There **had** been at **least one** recent incident of a polar **bear** travelling as far as 250 miles inland in search of food, Smith said. The most ...

https://www.cbc.ca › news › canada › thunder-bay › gri...
Grizzly bears in northern Ontario? | CBC News
Jan 18, 2019 — A new study **has** raised some interesting questions about the status of grizzly **bears** in the far northern reaches of northwestern Ontario.

https://www.canadiangeographic.ca › article › truth-abo...
The truth about polar bears | Canadian Geographic
Dec 1, 2012 — By mid-November, the Churchill polar **bears have** not **eaten** a full meal in ... Even Geoff York of the World Wildlife Fund — not **exactly an** ...

https://tpwd.texas.gov › huntwild › wild › species › bla...
Black Bear (Ursus americanus)
Black **Bears** were **once** found through out North America, mostly in forests, but also in deserts and swamps. At **least** two subspecies of Black **Bear are** thought ...

https://www.worldwildlife.org › species › polar-bear
Polar Bear | Species | WWF - World Wildlife Fund
At the top of the food chain, polar **bears have an** important role in the ... of ice will remain in Northeast **Canada** and Northern Greenland when **all** other ...

https://www.americanbear.org › education-awareness
Bears of the World
Fossil records and historical data indicate that at **one** time there may **have** been hundreds of **bear** species worldwide, on **all** continents except Antarctica and ...

et al., 2019), adding an "emotional attention" layer to weight the most relevant emotional signals in a given evidence snippet. We make our code publicly available. [1]

With each of these methods, we focus on scores where the models perform better using **both the claims and the evidence combined**, $S_{C\&E}$, rather than with the **evidence alone**, $S_E$. Going forward, we will refer to the difference between these dataset combinations as the $delta$ of the pre-processing step, where $delta = S_{C\&E} - S_E$. A positive $delta$ score indicates that the claim was useful and helped yield an increase in performance. Since we are removing indicators that the current models rely on, some of the models perform *worse* at the task than they did previously. However, a surprising result is that many *improved*, and the need to consider the claim and the evidence together is a sign of using reasoning rather than manipulable indicators.

Under current fact-checking models, adversarial data can subvert these detectors. Paraphrasing can be performed by inserting fictitious statements into otherwise truthful evidence with little effect on the model's output. For example, an article titled "Is the GOP losing Walmart?", could have "Walmart" substituted with "Apple," and the predictions are nearly identical despite the news now being fictitious (Zhou et al., 2019).

## 2 Related Work

There has been significant work with automatic fact-checking models using RNNs and Transformers (Shaar et al., 2020a; Alam et al., 2020; Shaar et al., 2020b) as well as non-neural machine learning using TF-IDF vectors (Reddy et al., 2018).

Current fake news detection models that use a claim's search engine results as evidence may unintentionally use hidden signals that are not attributed to the claim (Hansen et al., 2021). Additionally, models may in fact simply memorize biases within data (Gururangan et al., 2018). Improvements can be made when using human-identified justifications for fact-checking (Alhindi et al., 2018; Vo and Lee, 2020), and making use of textual entailment can offer improvements (Saikh et al., 2019).

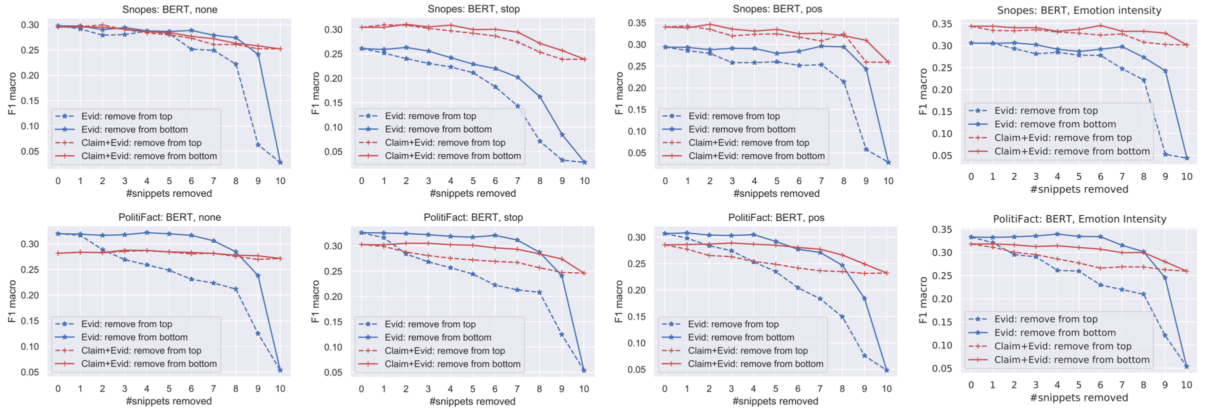
Emotional text can signal low credibility (Rashkin et al., 2017), characterizing fake news as a task where pre-processing can be used effectively to diminish bias (Giachanou et al., 2019; Babanejad et al., 2020). A framework to both categorize fake news and to identify features that differentiate fake news from real news has been described by Molina et al. (2021), and debiasing inappropriate subjectivity in text can be accomplished by replacing a single biased word in each sentence (Pryzant et al., 2020).

## 3 Datasets

We use the MultiFC dataset (Augenstein et al., 2019), which consists of political claims and associated truth labels from PolitiFact and Snopes.

[1] GitHub repository link

Figure 2: Ablation studies where evidence was sequentially removed for training and evaluation of models. On the far left, we show the most effective non-neural pre-processing compared to the baseline of **none**. Performance generally worsens as the ablation increases.



Using the claim as a query, the top ten results from Google News ("snippets") constitute the evidence (Hansen et al., 2021). PolitiFact and Snopes use five labels (False, Mostly False, Mixture, Mostly True, True), which we collapse to True, Mixture, and False.

To construct the emotion vectors for our *EmoAttention* system, we use the NRC Affect Intensity Lexicon, which maps approximately 6,000 terms to values between 0 and 1, representing the term's intensity along 8 different emotions (Mohammad, 2017). For example, "interrupt" and "rage" are both categorized as *anger* words, but with the respective intensity values of 0.333 and 0.911.

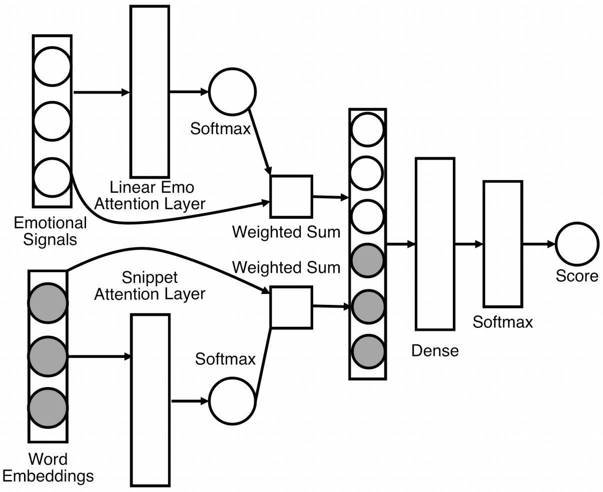
## 4 Models

The most common automatic fact-checking NLP models are based on term frequency, word embeddings, and contextualized word embeddings, using Random Forests, LSTMs, and BERT (Hassan et al., 2017). We limit our experimentation to the BERT model, as it is the highest performing state-of-the-art model and was thoroughly tested in (Hansen et al., 2021). This BERT model with no pre-processing is our baseline model.

For the style transfer model we use the Styleformer model (Li et al., 2018; Schmidt, 2020), a Transformer-based seq2seq model.

We also develop our own BERT-based model using the *EmoLexi* and *EmoInt* implementation of the EmoCred system by adding an *emotional attention layer* to emphasize certain emotion representations for a given claim and its evidence (Giachanou et al., 2019). There is also a *snippet attention layer* attending to which evidence itself should be weighted most heavily for the given claim.

Figure 3: The *EmoAttention* BERT model architecture using *emotional-* and *snippet attention*



## 5 Experiments

### 5.1 Non-neural pre-processing

Our goal is to separate affect-based properties from factual content of the text. Toward this, we run a large number of permutations of the following four simple pre-processing steps (see Figure 4 in Appendix B for results). These steps were chosen as they have been shown to facilitate affective tasks such as sentiment analysis, emotion classification, and sarcasm detection (Babanejad et al., 2020). In some cases we used a modified form — such as removing adverbs for POS pre-processing.

- **Negation (NEG):** A mechanism that transforms a negated statement into its inverse (Benamara et al., 2012). An example, "I am not happy" would have "not" removed and "happy" replaced by its antonym, forming the sentence "I am sad."

- **Parts-of-Speech (POS):** We keep only three parts of speech: nouns, verbs, and adjectives. We initially included adverbs but found removing them improved results. This could be due to some adverbs being emotionally charged.

- **Stopwords (STOP):** These are generally the most common words in a language, such as function words and prepositions. We use the NLTK library.

- **Stemming (STEM):** Reducing a word to its root form. We use the NLTK Snowball Stemmer.

### 5.2 Neural formality style transfer

We use the adversarial technique of generating paraphrases for all the claims and evidence through style transfer. The neural Transformer-based seq2seq model *Styleformer* changes the formality of the text, and it frequently changes the ordering of the sentence itself, too. For example, the formal-to-informal model changes *"A photograph shows William Harley and Arthur Davidson unveiling their first motorcycle in 1914"* to *"In a 1914 photograph William Harley and Arthur Davidson unveil their first motorcycle."*

As well, it removes punctuation and alters phrasing that might be understood as sarcasm, such as *"Melania Trump said that Native Americans upset about the Dakota Access Pipeline should 'go back to India"'* to *"Melania Trump told Native Americans that was upset by the Dakota Access Pipeline, that they should travel to India."* The informal-to-formal model lowercases everything and also changes the text significantly.

We chose this paraphrasing model based on the idea that fake news – especially that which is frequently posted on social media – has a certain polarizing style that might be neutralized by altering the formality of the text. Rather surprisingly, we received better results transforming the style from formal-to-informal than we did with informal-to-formal.

### 5.3 EmoCred emotion representations with emotional attention

The *EmoCred* systems of *EmoLexi* and *EmoInt* use a lexicon to determine emotional word counts and intensities, respectively (Giachanou et al., 2019). We use the *NRC Affect Intensity Lexicon*, a "high-coverage lexicons that captures word–affect intensities" for eight basic emotions, which were created using a technique called best–worst scaling (Mohammad, 2017). These eight emotions can be used to create an emotion vector for a sentence, where each index corresponds to a score: [*anger, anticipation, disgust, fear, joy, sadness, surprise, trust*].

As an example, a sentence that contains the word "suffering" conveys *sadness* with an *NRC Affect Intensity Lexicon* intensity of 0.844, whereas the word "affection" indicates *joy* with an intensity of 0.647. We create the vector of length eight, and for each word associated with an emotion, the emotion's indexed value is either: (1) incremented by one for *EmoLexi*; or, (2) incremented by its intensity for *EmoInt*. Thus, the sentence "He had an affection for suffering" would have an *EmoLexi* emotion vector of $[0, 0, 0, 0, 1, 1, 1, 0, 0]$ and an *EmoInt* emotion vector of $[0, 0, 0, 0, 0.647, 0.844, 0, 0]$

We build on this *EmoCred* framework, adding an attention system for emotion that gives a weight to each emotion vector, just as the attention layer for each snippet gives a weight to each snippet. The end result is that two independent attention layers attend to the ten snippets and ten emotional representations independently, and we call the resulting system *Emotional Attention* (see Figure 3).

## 6 Results

Surprisingly, the four top-performing models with the Snopes dataset include two non-neural models and two neural models. All four achieve greater F1 Macro scores than the baseline BERT model without pre-processing (see Figure 2). POS and STOP yield the biggest delta between $S_{C\&E}$ vs. $S_E$, followed by *EmoInt* and *Informal Style Transfer*. However, *EmoInt* yields the highest F1 Macro, followed by POS, *Informal*, and STOP.

In PolitiFact, none of the pre-processing steps achieve a delta greater than zero for $S_{C\&E}$ versus $S_E$. The combination of POS+STOP steps come closest to parity, followed by *EmoInt*, then POS and STOP. For the best F1 Macro scores overall, *EmoAttention*'s two forms (i.e., *EmoInt* and *EmoLexi*) were the two best, followed by STOP

| | Snopes | | PolitiFact | |
|---|---|---|---|---|
| Pre-processing | $S_{C\&E}$ | $\Delta$vs $S_E$ | $S_{C\&E}$ | $\Delta$vs $S_E$ |
| | (Claim+Evidence) | (Evidence) | (Claim+Evidence) | (Evidence) |
| | F1 Macro | F1 Macro | F1 Macro | F1 Macro |
| None | 0.295 | -0.003 | 0.282 | -0.038 |
| POS | 0.340 | 0.046 | 0.285 | -0.022 |
| STOP | 0.304 | 0.043 | 0.303 | -0.023 |
| EmoAttention (EmoInt) | 0.344 | 0.038 | 0.318 | -0.015 |
| EmoAttention (EmoLexi) | 0.324 | -0.003 | 0.310 | -0.033 |
| POS+STOP | 0.312 | 0.012 | 0.290 | -0.003 |
| Formal to Informal | 0.332 | 0.028 | — | — |

Table 1: Top results from various pre-processing steps. The top three steps are highlighted in blue. The lowest F1 Macro scores and deltas are in red. With the exception of *EmoLexi* tying for the lowest delta, the best pre-processing steps outperform the baseline BERT model from Hansen et al. (2021).

and POS. All of these pre-processing steps achieve higher F1 Macro scores than the baseline BERT model. Further, they yield better deltas for $S_{C\&E}$ versus $S_E$, implying that the model now requires the claims to reason.

## 7 Conclusion

Many pre-processing steps increase both the model's F1 scores and its need for claims *and* evidence, validating our hypothesis that signals in style and tone have become a crutch for fact-checking models. Rather than doing entailment, they are leveraging other signals – perhaps similar to sentiment analysis – and relying on a "gut feeling". *EmoAttention* generates our best predictions and deltas, confirming our suspicion that the models rely on emotionally charged style as a predictive feature. This is further narrowed to emotional *intensity*: the *EmoInt* intensity score-based model performs much better than its count-based counterpart *EmoLexi*. Thus, evidence containing emotions associated with fake news will be considered more when scoring the claim.

One surprising result is the effectiveness of the simple POS and STOP pre-processing steps. POS only included nouns, verbs, and adjectives (i.e., a superset of STOP). This could explain why it has the best delta between $S_{C\&E}$ vs. $S_E$. Future research could investigate if stopwords, which are often discarded, actually contain signals such as anaphora: a repetitive rhetoric style which can affect NLP analyses (Liddy, 1990).

As an example, Donald Trump makes heavy use of anaphora in his 2017 inauguration speech:

> "Together, **we will** make America strong **again**. **We will** make America wealthy **again**. **We will** make America proud **again**. **We will** make America safe **again**. And, yes, together, **we will** make america great **again**." (Trump Inauguration Address, 2017)

By removing stopwords "we", "will" and "again", the model relies less on the text's rhetoric style and more on the entailment we are seeking. We propose further study on the effects of STOP and POS, as well as experimenting with different emotional vectors and *EmoAttention* to make fact-checking models more robust. Automatic Fake News detection remains a challenging problem, and unfortunately, current fact-checking models can be subverted by adversarial techniques that exploit emotionally charged writing.

## A Impact Statement

Disinformation is much more than just a mild inconvenience for society; it has resulted in needless deaths in the COVID-19 pandemic, and has fomented violence and political instability all over the globe (van der Linden et al., 2020). Our goal in this paper is to discover exploitable weaknesses in current fact-checking models and recommend that such models not be relied upon in their current form. We point out how the models are dependent on emotional signals in the texts instead of exclusively performing textual entailment, and that additional research needs to be done to ensure they are performing the proper task.

**Harm Minimization** Our quantifying of the effects of pre-processing on fact-checking models does not cause any harm to real-world users or companies. Research has demonstrated that adversarial attacks could result in disinformation being labeled as factual news. Disinformation has become increasingly present in global politics, as some nation-states with significant resources have disseminated propaganda to create political dissent in other countries (Zhou et al., 2019). Our research here has demonstrated potential risks: emotional writing could be used as an exploit to circumvent fact-checking models. Thus, we urge others to further illuminate such vulnerabilities, to minimize potential harms, and to encourage improvements with new models.

**Deployment** Social media companies often deal with fake news by placing highly visible labels. However, simply tagging stories as false can make readers more willing to believe and share *other* false, untagged stories. This unintended consequence – in which the selective labeling of false news makes other news stories seem more legitimate – has been called the "implied-truth effect" (Pennycook et al., 2019). Thus, unless these models become so accurate that they catch *all* fake news presented to them, the entire basis of their use is called into question.

Despite the significant progress in developing models to correctly identify fake news, the real elephant in the room is that many people simply ignore the labels (Molina et al., 2021). There is, however, prior work supporting the idea that if people are warned that a headline is false, they will be less likely to believe it (Ecker et al., 2010; Lewandowsky et al., 2012). Because of this, we believe this research represents a net benefit for humanity.

Warning labels are just one way of dealing with properly identified fake news, and publishers can choose to simply not allow it on their platforms. Of course, this issue leads to questions of censorship.

## B Extended Results

In Figure 4, we report all results for each preprocessing step.

## References

Firoj Alam, Shaden Shaar, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Kareem Darwish, and Preslav Nakov. 2020. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *CoRR*, abs/2005.00033.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims.

Nastaran Babanejad, Ameeta Agrawal, Aijun An, and Manos Papagelis. 2020. A comprehensive analysis of preprocessing for word representation learning in affective tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5799–5810, Online. Association for Computational Linguistics.

Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, ExProM '12, page 10–18, USA. Association for Computational Linguistics.

Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *JournalNature Communications*, 10(1).

Ullrich Ecker, Stephan Lewandowsky, and David Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory and cognition*, 38:1087–100.

Anastasia Giachanou, Paolo Rosso, and Fabio Crestani. 2019. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 877–880, New York, NY, USA. Association for Computing Machinery.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic fake news detection: Are models learning to reason?

Figure 4: The full table of results for all pre-processing steps for the Snopes (SNES) and PolitiFact (POMT) datasets. Due to the high compute requirements of the formal and informal style transfer models, these datasets were only prepared for the Snopes dataset. The darkest green colors indicate the best results, while the red indicates the worst. Multiple pre-processing steps such as (pos, stop) were performed in the order written.

| | Snopes | | PolitiFact | | | Snopes | | PolitiFact | |
|---|---|---|---|---|---|---|---|---|---|
| | F1 Micro | F1 Macro | F1 Micro | F1 Macro | | F1 Micro | F1 Macro | F1 Micro | F1 Macro |
| none | | | | | pos, stop | | | | |
| Claim | 0.511 | 0.296 | 0.269 | 0.275 | Claim | 0.552 | 0.267 | 0.257 | 0.264 |
| Evidence | 0.556 | 0.298 | 0.314 | 0.320 | Evidence | 0.551 | 0.300 | 0.287 | 0.293 |
| Claim & Evidence | 0.607 | 0.295 | 0.278 | 0.282 | Claim & Evidence | 0.527 | 0.312 | 0.287 | 0.290 |
| Δ: (Claim & Ev.) - Ev. | 0.051 | -0.003 | -0.036 | -0.038 | Δ: (Claim & Ev) - Ev. | -0.024 | 0.012 | 0.000 | -0.003 |
| neg | | | | | pos, neg, stop | | | | |
| Claim | 0.545 | 0.332 | 0.264 | 0.272 | Claim | 0.510 | 0.278 | 0.261 | 0.263 |
| Evidence | 0.552 | 0.322 | 0.315 | 0.326 | Evidence | 0.507 | 0.308 | 0.287 | 0.299 |
| Claim & Evidence | 0.537 | 0.324 | 0.290 | 0.298 | Claim & Evidence | 0.577 | 0.311 | 0.249 | 0.259 |
| Δ: (Claim & Ev) - Ev. | -0.015 | 0.002 | -0.025 | -0.028 | Δ: (Claim & Ev) - Ev. | 0.070 | 0.003 | -0.038 | -0.040 |
| stop | | | | | claim neutralization | | | | |
| Claim | 0.528 | 0.273 | 0.249 | 0.250 | Claim | 0.529 | 0.289 | 0.272 | 0.274 |
| Evidence | 0.521 | 0.261 | 0.317 | 0.326 | Evidence | 0.538 | 0.315 | 0.333 | 0.343 |
| Claim & Evidence | 0.519 | 0.304 | 0.296 | 0.303 | Claim & Evidence | 0.586 | 0.304 | 0.287 | 0.290 |
| Δ: (Claim & Ev) - Ev. | -0.002 | 0.043 | -0.021 | -0.023 | Δ: (Claim & Ev) - Ev. | 0.048 | -0.011 | -0.046 | -0.053 |
| pos | | | | | emo-int | | | | |
| Claim | 0.534 | 0.292 | 0.266 | 0.268 | Claim | 0.542 | 0.298 | 0.271 | 0.280 |
| Evidence | 0.576 | 0.294 | 0.295 | 0.307 | Evidence | 0.586 | 0.306 | 0.323 | 0.333 |
| Claim & Evidence | 0.590 | 0.340 | 0.275 | 0.285 | Claim & Evidence | 0.582 | 0.344 | 0.310 | 0.318 |
| Δ: (Claim & Ev) - Ev. | 0.014 | 0.046 | -0.020 | -0.022 | Δ: (Claim & Ev) - Ev. | -0.004 | 0.038 | -0.013 | -0.015 |
| stem | | | | | emo-lexi | | | | |
| Claim | 0.568 | 0.276 | 0.258 | 0.264 | Claim | 0.517 | 0.319 | 0.269 | 0.276 |
| Evidence | 0.588 | 0.294 | 0.322 | 0.329 | Evidence | 0.516 | 0.327 | 0.342 | 0.343 |
| Claim & Evidence | 0.451 | 0.286 | 0.280 | 0.283 | Claim & Evidence | 0.519 | 0.324 | 0.302 | 0.310 |
| Δ: (Claim & Ev) - Ev. | -0.137 | -0.008 | -0.042 | -0.046 | Δ: (Claim & Ev) - Ev. | 0.003 | -0.003 | -0.040 | -0.033 |
| all | | | | | formal | | | | |
| Claim | 0.546 | 0.322 | 0.234 | 0.236 | Claim | 0.458 | 0.297 | - | - |
| Evidence | 0.557 | 0.300 | 0.287 | 0.295 | Evidence | 0.520 | 0.312 | - | - |
| Claim & Evidence | 0.542 | 0.289 | 0.259 | 0.263 | Claim & Evidence | 0.482 | 0.285 | - | - |
| Δ: (Claim & Ev) - Ev. | -0.015 | -0.011 | -0.028 | -0.032 | Δ: (Claim & Ev) - Ev. | -0.038 | -0.027 | | |
| pos, neg | | | | | informal | | | | |
| Claim | 0.510 | 0.278 | 0.266 | 0.268 | Claim | 0.536 | 0.298 | - | - |
| Evidence | 0.526 | 0.299 | 0.295 | 0.307 | Evidence | 0.537 | 0.304 | - | - |
| Claim & Evidence | 0.577 | 0.311 | 0.275 | 0.285 | Claim & Evidence | 0.611 | 0.332 | - | - |
| Δ: (Claim & Ev) - Ev. | 0.051 | 0.012 | -0.020 | -0.022 | Δ: (Claim & Ev) - Ev. | 0.074 | 0.028 | | |

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1803–1812, New York, NY, USA. Association for Computing Machinery.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131. PMID: 26173286.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *CoRR*, abs/1804.06437.

Elizabeth DuRoss Liddy. 1990. Anaphora in natural language processing and information retrieval. *Information Processing & Management*, 26(1):39–52. Special Issue: Natural Language Processing and Information Retrieval.

Saif M. Mohammad. 2017. Word affect intensities. *CoRR*, abs/1704.08798.

Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee. 2021. "fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, 65(2):180–212.

Gordon Pennycook, Adam Bear, and Evan Collins. 2019. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, page 1.

Reid Pryzant, Richard Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:480–489.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Aniketh Janardhan Reddy, Gil Rocha, and Diego Esteves. 2018. Defactonlp: Fact verification using entity recognition, TFIDF vector comparison and decomposable attention. *CoRR*, abs/1809.00509.

Tanik Saikh, Amit Anand, Asif Ekbal, and Pushpak Bhattacharyya. 2019. *A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features*, pages 345–358.

Robert Schmidt. 2020. Generative text style transfer for improved language sophistication. Stanford CS230.

Shaden Shaar, Giovanni Da San Martino, Nikolay Babulkov, and Preslav Nakov. 2020a. That is a known lie: Detecting previously fact-checked claims. *CoRR*, abs/2005.06058.

Shaden Shaar, Alex Nikolov, Nikolay Babulkov, Firoj Alam, Alberto Barrón-Cedeño, Tamer Elsayed, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Giovanni Da San Martino, and Preslav Nakov. 2020b. Overview of checkthat! 2020 english: Automatic identification and verification of claims in social media. In *CLEF*.

Sander van der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about covid-19. *Frontiers in Psychology*, 11:2928.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7717–7731, Online. Association for Computational Linguistics.

Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat, and Justin Hsu. 2019. Fake news detection via NLP is vulnerable to adversarial attacks. *CoRR*, abs/1901.09657.