Accenture at CheckThat! 2023: Learning to Detect Factuality Levels of News Sources

Sieu Tran[1], Paul Rodrigues[1], Benjamin Strauss[1] and Evan M. Williams[2]

[1]*Accenture, 1201 New York Ave NW, Washington, DC 20005, United States*
[2]*Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States*

Abstract

News sources are expected to present factual evidence with accurate and unbiased reasoning to inform their readership. A news source that repeatedly presents reporting with low factuality can erode confidence and trust and lose a reader. With a low barrier to technical entry, numerous news sources have emerged, and it can be difficult to evaluate a new source for accuracy, especially if one is not familiar with their reputation. Tools such as social media and news aggregators can utilize social signals, but the social network propagating the news may not be motivated or qualified to evaluate the source for factuality. Propagation of false and misleading information through these networks can harm decision making of a group and cause spread of misinformation. This paper presents a model that learns the factuality levels of news articles and news sources, using an NLP data augmentation strategy to increase the size of the training data. The model can then be applied to unseen news sources and articles to identify the factuality level of the document. A system, like this one, could be utilized in a news aggregator or social network to flag an article with a credibility warning or to deactivate user interface elements that promote article dissemination.

Keywords

fact detection, fact checking, factuality, factivity

1. Introduction

As news information ecosystems become increasingly fractured and divisive, and with the increase of citizen journalism causing an increase in news sources, users must constantly evaluate the veracity of the sources to which they are exposed. Automated reliability detection systems can remove some of this burden from the user and can aid search engines, news aggregators, and social media recommendation systems in leading users to accurate and reliable content. Our submission to CLEF's CheckThat! 2023 Task 4 develops an automated news source reliability detection system.

Detecting reliability based on a subset of articles for each domain is an inherently challenging task as the lines between "high", "mixed", and "low" reliability are often blurry. It can be unclear where cut-offs between each category should lie. Authors of articles can insinuate false conspiracies or explanations, recount stories in misleading and biased ways, or ask leading

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

sieu.tran@accenturefederal.com (S. Tran); paul.rodrigues@accenturefederal.com (P. Rodrigues); b.strauss@accenturefederal.com (B. Strauss); emwillia@andrew.cmu.edu (E. M. Williams)
0000-0003-0017-4329 (S. Tran); 0000-0002-2151-636X (P. Rodrigues); 0000-0002-0224-424X (B. Strauss); 0000-0002-0534-9450 (E. M. Williams)

questions all without ever stating a technical falsehood. Even on highly-unreliable sites, some articles often contain true or mostly-true information. For example, In the task 4 data provided by organizers, "dcclothesline", an unreliable news domain, has an article that refers to the COVID-19 vaccine as a "genocide", but also has an article that accurately, albeit with biased language, recounts a story about Rashida Tlaib falsely attributing the death of a Palestinian child to "Israeli Settlers" on Twitter [1].

Several works have used web graph data and search engine optimization data to explore statistical connections between unreliable news domains as well as the approaches the domains take to manipulate search engines [2, 3, 4]. While reliability detection remains rare in the literature, other misinformation detection tasks often leverage domain-level data. [5] jointly predict news media reliability as well as bias using Copula Ordinal Regression (COR) models. Many researchers use site reliability labels within a larger detection system to detect false or misleading news (e.g. [6, 7])

In this work, we describe the data augmentation and fine tuning approach employed by the Accenture Team for CLEF CheckThat! Lab's Task 4. Teams were tasked with classifying websites sources as "low", "mixed", or "high" factuality from their articles and evaluated using mean absolute error (MAE).

2. Exploratory Analysis

Table 1 shows the number of samples and unique word counts for each of the datasets provided.

Table 1
Dataset descriptions

Task	Modeling Group	# of Source	# of Articles	Unique Words
Factuality News Media Source	Train	947	7,948	98,725
Factuality News Media Source	Test	122	1,054	36,766
Factuality News Media Source	Validation	120	1,049	36,157

2.1. Label Balance

The training dataset had label bias which skewed towards sources that were "high" labeled on factuality: 61% high, 26% mixed, and 13% low.

2.2. WordPiece Analysis

Transformer models utilize WordPiece tokenization schemes that are dependant on the model being evaluated. At the time of pre-training, the WordPiece algorithm determines which pieces of words will be retained, and which will be discarded. We present our analysis in Table 2. Unexpectedly, the RoBERTa tokenizers we used did not return UNK tokens on any dataset provided by the CLEF CheckThat! organizers.

Table 2
Token distribution in data

Tokenizer Type	Modeling Set	WordPiece
RoBERTa-based	Train	8,645,632
	Test	1,103,855
	Validation	1,057,250

3. Transformer Architectures and Pre-Trained Models

In this work, we utilize RoBERTa models. The Bidirectional Encoder Representation Transformer (BERT) is a transformer-based architecture that was introduced in 2018 [8, 9]. BERT has had a substantial impact on the field of NLP, and achieved state of the art results on 11 NLP benchmarks at the time of its release. RoBERTa, introduced by [10], modified various parts of BERTs training process. These modifications include more training data, more pre-training steps with bigger batches over more data, removing BERT's Next Sentence Prediction, training on longer sequences, and dynamically changing the masking pattern applied to the training data [11]. For this work, we fine-tune *roberta-large* [12]. The English RoBERTa model contains 50,265 WordPieces.

4. Method

4.1. Data Augmentation

The organizers provided a training and a development set for each language. We use the provided training set and development set to create internal training and validation sets for experimentation. We use the test set provided by organizers as a hold-out test set.

For each article, augmentation and training were done with via back-translation using AWS translation. We appended back-translated low and mixed factuality articles to the training set. In our 2021 experiment [13], we found that this form of augmentation resulted in a significant increase in recall and F1-score for check-worthy tweets. Due to significant sample imbalance in the training sets for both task, we augmented the 0 (Low)- and 1 (Mixed)-class until the samples are roughly balanced. Table 3 shows the BLEU score for each back-translation scheme performed. BLEU score is historically used to compute the quality of machine translation by comparing machine translated text to a human translated reference. When using BLEU score for data augmentation the lower the BLEU score, the more divergent the translation to the original text, providing more diverse training data.

The number of new tokens added to the text corpus by back translation can be found in 4. Adding a large number of useful new tokens aids in diversifying the model.

4.2. Classification

For the RoBERTa model, we added an additional mean-pooling layer and dropout layer on top of the model prior to the final three binary classification layers, each of which corresponding to

Table 3
Average Sentence BLEU Score for Each Back-translation Scheme

Label	Back-translation	Avg Sentence BLEU Score
0 (Low)	EN > ES > EN	0.454
0 (Low)	EN > EN > EN > FR > EN	0.383
0 (Low)	EN > EN > EN > FR > EN > DE > EN	0.335
1 (Mixed)	EN > EN > EN	0.481

Table 4
New Tokens in Machine Translated Text

Label	Back-translation	Unique tokens in source	Unique tokens in MT	New Tokens in MT
0 (Low)	EN > SP > EN	58770	58156	18626
0 (Low)	EN > SP > EN > FR > EN	58770	57322	20356
0 (Low)	EN > SP > EN > FR > EN > DE > EN	58770	55945	21906
1 (Mixed)	EN > SP > EN	102970	93502	33677

a class (i.e., 0 (Low), 1 (Mixed), or 2 (High)). The highest class probability determines the article final classification. Adding these additional layers has been shown to help prevent over-fitting while fine-tuning. We used an Adam optimizer with a learning rate of $2e - 5$ and an epsilon of $1.5e - 8$. We use a binary cross-entropy loss function, 4 epochs, and a batch size of 32. The majority class of all articles released by a given news source determines the level at which the news source is factual.

5. Results

Table 5 shows model performance on the test set provided by the organizers. The classification task reached a weighted average F1-score of 0.592. The accuracy of the classifier was 0.590. This method received a Mean Absolute Error (MAE) of 0.467 in the official evaluation.

6. Conclusion

This paper introduced the methods and results from the Accenture Team for the 2023 CLEF CheckThat! Lab's Task 4, identifying the factuality of news sources. Teams were tasked with classifying websites sources as "Low", "Mixed", or "High" factuality from their article content and the tasks were evaluated using mean absolute error (MAE).

References

[1] S. J. Frantzman, Rashida Tlaib retweets unverified claim Israelis killed Palestinian boy (2020).

Table 5
Accenture results from CheckThat! 2023 Task 4

Class	Precision	Recall	F1-score
0 (Low)	0.500	0.263	0.345
1 (Mixed)	0.364	0.516	0.427
2 (High)	0.750	0.708	0.729
macro avg	0.538	0.496	0.500
weighted avg	0.614	0.590	0.592

Accuracy	MAE
0.590	0.467

[2] V. Mazzeo, A. Rapisarda, Investigating fake and reliable news sources using complex networks analysis, Frontiers in Physics 10 (2022) 886544.

[3] E. M. Williams, K. M. Carley, Search engine manipulation to spread pro-kremlin propaganda, Harvard Kennedy School Misinformation Review (2023).

[4] S. Bradshaw, Disinformation optimised: gaming search engine algorithms to amplify junk news, Internet policy review 8 (2019) 1–24.

[5] R. Baly, G. Karadzhov, A. Saleh, J. Glass, P. Nakov, Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media, arXiv preprint arXiv:1904.00542 (2019).

[6] B. Ghanem, S. P. Ponzetto, P. Rosso, F. Rangel, Fakeflow: Fake news detection by modeling the flow of affective information, arXiv preprint arXiv:2101.09810 (2021).

[7] I. Baris, Z. Boukhers, Ecol: E arly det ec tion of co vid l ies using content, prior knowledge and source information, in: Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers 1, Springer, 2021, pp. 141–152.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[9] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, 2019. arXiv:1908.08962.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[11] E. M. Williams, P. Rodrigues, V. Novak, Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of claims using transformer-based models, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_226.pdf.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoy-

anov, RoBERTa: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[13] E. Williams, P. Rodrigues, S. Tran, Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, 2021. arXiv:2107.05684.