Identifying Pseudo-science via Deep Bidirectional Language Models and Computational Linguistics

Or Levi, Inbal Croitoru, Reenah Nahum
{or,inbal,reenah}@adverifai.com
AdVerif.AI

ABSTRACT

Pseudo-science, a collection of beliefs or practices mistakenly regarded as being based on scientific method, can have harmful implications including the loss of life or health as recently evidenced during the the COVID-19 pandemic. While a great body of work has been dedicated in recent years for identifying fake news, research on pseudo-scientific content has remained largely unexplored presumably due to the lack of publicly available data. In this work, we propose a novel large-scale, well-balanced and source-rich dataset to support research on detection of harmful pseudo-scientific content. Alongside the dataset, we provide several strong baselines, using fine-tuned deep bidirectional language models, computational linguistics features and a combination of thereof. Empirical evaluation attests to the merits of our approach for effectively identifying pseudo-science, and demonstrates the efficacy of the linguistic features, with the combined method achieving a significantly higher accuracy than each method separately.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Machine learning**;

KEYWORDS

misinformation, transformers, computational linguistics

ACM Reference Format:
Or Levi, Inbal Croitoru, Reenah Nahum. 2021. Identifying Pseudo-science via Deep Bidirectional Language Models and Computational Linguistics. In *KDD MIS2 '21: Misinformation and Misbehavior Mining on the Web, August 14–18, 2021.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Pseudo-science consists of statements, beliefs, or practices that are claimed to be both scientific and factual but are incompatible with the scientific method [1]. Such claims, that often reject modern medicine in favor of alternative medicine, have led on numerous occasions to loss of life or health. In the last year, the COVID-19 pandemic has reinforced the potential of health-related misinformation and pseudo-science to cause vast economic and health damages[1]. It was estimated that more than 70% of the UK Citizens consumed misinformation online with more than 2 million people who have changed their behavior as a result. This had an indirect impact on the UK economy of £3.6 billion in the second and third quarters of 2020, and caused more than 500 additional deaths.

According to Lilienfeld et al. [7], pseudo-science differs from science in multiple aspects, including lack of self-correction, lack of falsifiability, emphasis on the confirmation, evasion of peer review, over-reliance on testimonial and anecdotal evidence, and more. Nonetheless, pseudo-science often preys on cognitive biases and can be challenging to decipher by the untrained observer. In the past few years, a great body of work has been dedicated to studying the characteristics of fake news content and developing methods for automatically identifying fake news. Compared to that, and despite the growing interest in identifying harmful pseudo-scientific content, it has remained largely unexplored presumably due to the lack of publicly available data.

In this work, we propose a novel large-scale dataset to advance research on pseudo-science detection. The dataset holds three desirable proprieties, that are 1) including a multitude of diverse pro-science and pseudo-science sources; 2) providing a well-balanced distribution of articles across sources; and 3) effectively measuring generalization and supporting the development of robust models.

Furthermore, alongside the dataset, we present methods for identifying pseudo-science content, that combine transformer-based language models and computational linguistics. While previous work[8, 9] have demonstrated that pro-science and pseudo-science articles can be distinguished based on the language used and topical cues, our work also considers linguistic and grammatical aspects of writing. To this end, we study a variety of linguistic features, including lexical features, text readability indicators and grammatical tagging. Our main research question is therefore, can these linguistic and grammatical differences contribute to the understanding of nuances between pro-science and pseudo-science beyond differences in the language being used?

Overall, our contributions can be summarized as follows:

- We propose a novel large-scale dataset with a multitude of diverse sources to support the research on detection of harmful pseudo-scientific content.
- We present a new method using an ensemble of fine-tuned state-of-the-art transformers and computational linguistics to distinguish between pseudo-science and pro-science.
- Through empirical evaluation, we show that our approach is effectively identifying pseudo-science, with the combined

KDD MIS2 '21, August 14–18, 2021,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/1122445.1122456

[1] https://londoneconomics.co.uk/wp-content/uploads/2021/01/The-Cost-of-Lies_clean_2.2.21.pdf

method achieving a significantly higher accuracy than each method in isolation.

2 RELATED WORK

Several research studies [2] have conducted experiments to measure the ability of humans to detect fake news. To understand reader susceptibility, Perez et al.[10] shared celebrity-oriented real and fake news with humans, who achieved an average accuracy of 70.5% in detecting made-up crowd-sourced news, and 78.5% in detecting real news. This and similar studies have shown that humans can easily be deceived into believing false information, giving rise to the challenge of automatically identifying fake news.

• **Fake News Detection**. Previous work [3, 6] demonstrated the merits of transformer-based language models and computational linguistics for identifying deception and misinformation. Reis et al.[6] presented a variety of linguistic features for identifying fake news including language features, lexical features, psycho-linguistic features, semantic features and part-of-speech (POS) tagging. Levi et al. [3] investigated the semantic and linguistic differences of satire and fake news articles. They utilized BERT (Bidirectional Encoder Representations from Transformers) for semantic textual representation and extracted linguistic features based on textual coherence metrics.

• **Pseudo-Science Detection**. Shvets et al.[8] presented a method for identifying pseudo-scientific publications based on automatic text analysis. They observed that pseudo-scientific publications contain specific lexical and stylistic features. As classification features they used individual words, word-combinations with syntactic dependencies and tri-grams. Using a dataset of 34 pseudo-scientific and 288 scientific publications, they calculated statistical word representations and trained a Support Vector Machine (SVM) classifier.

Rajapakse et al.[9] introduced a dataset of 112,720 full-text articles and proposed a method to recognize pseudo-science in the text by utilizing a fine-tuned Robustly Optimized Bidirectional Encoder Representation from Transformers Approach (RoBERTa)[4]. The dataset introduced by [9] consists of 20 websites that were labeled as pseudo-science or pro-science through manual inspection of the site content, as well as referencing public curated lists. In their work, 83% of the articles were collected from 7 out of the 20 sites. Compared to that, our work proposes a novel dataset consisting of 200 sites, which is well-balanced and distributed such that each site contributes no more than 0.75% of all the articles. The dataset we propose aims to provide a diverse variety of pro-science and pseudo-science narratives to support the development of robust detection models.

3 DATASET

To remedy the lack of large-scale diverse data for studying pseudo-science content, we propose a novel dataset[2], containing 10,000 articles obtained from 100 pseudo-science and 100 pro-science sites. Sites in the dataset are classified as pseudo-science or pro-science based on the information available on the *Media Bias Fact Check (MBFC)* website[3], which contains manual annotations and analysis of the factuality of reporting and/or bias for over 2,000 news websites. According to MBFC, sites labeled as pseudo-science may publish unverifiable information that is not always supported by evidence, while pro-science sites consist of legitimate science and evidence based on the use of credible scientific sourcing. For each site in the data we collect a multitude of articles using the Python Newspaper toolkit and examine the linguistic and semantic properties of the two classes. Table 1 shows the top 10 most frequent pseudo-science and pro-science sites in our data. We include a word cloud of the most frequent keywords of the pseudo-science and pro-science classes shown in Figure 1.

Rank	Pseudo-science Sites	Science Sites
1	davidicke.com	aaas.org
2	judithcurry.com	bio.org
3	medicine.news	cosmosmagazine.com
4	prisonplanet.com	discovermagazine.com
5	americanfreepress.net	everydayhealth.com
6	ancient-code.com	futurism.com
7	anonhq.com	nasa.gov
8	australiannationalreview.com	phys.org
9	charismanews.com	sciencealert.com
10	creation.com	sciencemag.com

Table 1: Top 10 most frequent pseudo-science and science sites in the dataset

The proposed dataset holds 3 desirable properties:

• **Diversity of Sources**. A diverse list of sites is important for studying the variety of pseudo-science and pro-science narratives and for developing robust detection models. Through the information provided by MBFC, the number of unique sites in the dataset is 10 times larger than previously available data.

• **Well-Balanced Articles Distribution**. To generate a well-balanced representation of the sources we employ stratified sampling. For sites with less than 75 articles we consider all available articles, while for sites with more than 75 articles we sample a random subset of 75 articles. The threshold of 75 is selected such that the sites with a plethora of articles will complement the sites with a shortage of articles to have an average of 50 articles per site and 10,000 articles in total. As a result, each site contributes no more than 0.75% of the number of all articles.

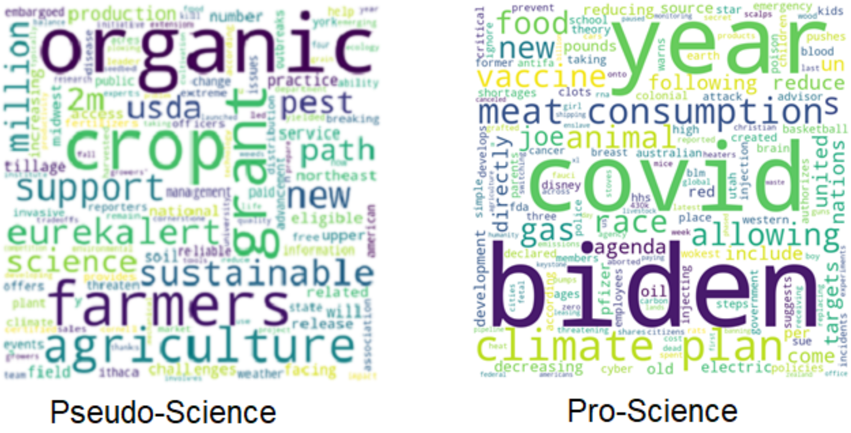
• **Robustness and Generalization**. We split the data with 80% and 20% of the articles assigned to train and held-out test sets, respectively, similar to [9]. However, to prevent over-fitting on patterns of specific sites, we separate the data such that each site is either in the train or the test set. We randomly sample 20% of the sites and the articles of these sites are used for the test set. Hence, the test set consists only of sites previously unseen by the model.

4 METHOD

In this section we describe the details of our method for identifying pseudo-science. First, we present the methods based on deep bidirectional language models which we fine-tune on the pseudo-science dataset. Next, we describe a variety of linguistic features related to lexical analysis, readability and grammatical tagging. Finally, we present an ensemble learning method combining both the fine-tuned language models and linguistic features.

[2] https://github.com/adverifai/pseudoscience_transformers

[3] https://mediabiasfactcheck.com



Pseudo-Science

Pro-Science

Figure 1: Most frequent keywords of the pseudo-science and pro-science classes

4.1 Deep Bidirectional Language Models

• **Fine-tuned Sentence-BERT**. To identify pseudo-science we utilize Sentence-BERT [5] for semantic representation of the textual articles. Sentence-BERT (SBERT) was presented in 2019 as a state-of-the-art sentence embeddings method that uses Siamese BERT-Networks. The model produces sentence embeddings with semantic meaning that allow textual relevance to be calculated using cosine similarity such that similar sentences are close to each other in the vector space. Thereby, SBERT is able to substantially shorten the computation time needed compared to BERT. We use SBERT as implemented in the Sentence Transformers repository[4] and leverage the pseudo-science dataset described in Section 3 to fine-tune the pre-trained *paraphrase-distilroberta-base-v2* version of SBERT.

• **Fine-tuned RoBERTa**. RoBERTa (Robustly Optimized Bidirectional Encoder Representation from Transformers Approach) was introduced by Liu et al [4] and achieved state-of-the-art results on the NLP multi-task General Language Understanding Evaluation (GLUE) benchmark. RoBERTa builds upon BERT (Bidirectional Encoder Representations from Transformers) with several modifications for generating a more robust model, specifically (1) training the model longer, with bigger batches, over more data; (2) removing the next sentence prediction objective; (3) training on longer sequences; and (4) dynamically changing the masking pattern applied to the training data. RoBERTa and BERT are pre-training a transformer model on a large-scale corpus which can be leveraged for effective learning in downstream tasks. These models are well-suited for use in text classification as the pre-trained model can be utilized with relatively small task-specific data through transfer learning. In this work, we fine-tune the pseudo-science detection model via transfer learning on a pre-trained RoBERTa model consisting of 12-layers of 768-hidden units, each with 12 attention heads, while leveraging the Hugging Face Transformers library[11]. We train the model using the pseudo-science dataset presented in Section 3. We consider the text of each article up to the maximum length limit of RoBERTa and follow the hyper-parameter recommendations by Liu et al. [4].

4.2 Linguistic Features

Inspired by previous work [3, 6] that demonstrated the effectiveness of computational linguistics for identifying fake news, we study a variety of linguistic features for identifying pseudo-science. The features are utilized for training a Logistic Regression model using the pseudo-science dataset. Overall, we evaluate a dozen of textual features, that are grouped in the following three sets:

• **Lexical**. The lexical features in our analysis include word-level and sentence-level signals. Word-level features are calculated by counting the amount of words in the text and the average number of syllables. To measure text complexity, we count the number of difficult words in the text, which are defined as words with more than three syllables and excluding the common words list defined by the TextStat library. For sentence-level features, we use the number of sentences and their average length.

• **POS Tagging**. We use part-of-speech (POS) tagging as implemented by the spaCy NLP toolkit. The part-of-speech of each sentence used in our model is encoded with a TF-IDF (Term Frequency – Inverse Document Frequency) representation.

• **Readability**. To evaluate writing style as a potential indicator of text quality, we extract features related to text readability. To measure the readability we use three formulas, including the Gunning Fog Index score calculated by:

$$0.4 * [(words/sentences) + 100 * (complex\ words/words)]; \quad (1)$$

The SMOG grading formula, that estimates the years of education needed to understand a piece of writing, given by:

$$1.0430 * \sqrt{(}polysyllables * 30/sentences) + 3.1291; \quad (2)$$

and the Dale–Chall readability formula, which also indicates the comprehension difficulty of text:

$$0.1579*(difficultwords/words*100)+0.0496*(words/sentences). \quad (3)$$

4.3 Ensemble Learning

Using the fine-tuned SBERT and RoBERTa language models and the linguistic features, we train an ensemble model to combine the predictions for identifying pseudo-science. We further split the

[4] https://github.com/UKPLab/sentence-transformers

Model	Accuracy	F1 Science	F1 Pseudo-science
Linguistic Features	71.70	74.32	68.48
Sentence-BERT	81.29	82.31	80.14
Sentence-BERT + Linguistic Features	82.17	83.01	81.26
RoBERTa	88.29	89.28	87.13
RoBERTa + Linguistic Features	**89.51**	**90.35**	**88.47**

Table 2: Main Results. Results of classification between pseudo-science and pro-science articles using linguistic features, Sentence-BERT, RoBERTa and ensemble learning methods. Bold: best performing model.

Feature Group	Accuracy	F1 Science	F1 Pseudo-science
All Linguistic Features	71.70	74.32	68.48
-Readability	71.36	74.15	67.88
-Lexical	67.82	70.08	65.19
-POS Tagging	61.08	66.51	53.57

Table 3: Linguistic Features Importance. On each row, a single group of features is removed from the model.

train set into 80% and 20% for training and validation, respectively. First, we train each of the SBERT and RoBERTa language models using the train set. Then, we use the validation set to train a Logistic Regression model using the predictions of each of the language models and the linguistic features to produce a final prediction.

5 EVALUATION

We use the dataset of pseudo-science and pro-science articles described in Section 3 to evaluate the performance of our method. The predictive performance is measured using the F1 score, balancing both precision and recall.

• **Main Results**. Table 2 shows the main results of our evaluation. We compare the performance of the fine-tuned Sentence-BERT model, the fine-tuned RoBERTa model and the linguistic features, as well as the combined models using both language models and the linguistic features, as described in Section 4. The method using only the linguistic features demonstrates the utility of these features for effectively identifying pseudo-science. It can be observed that the models based on Sentence-BERT perform relatively worse compared to the RoBERTa-based models in identifying pseudo-science. However, the addition of the linguistic features significantly improves the performance for both models. The two-tailed paired t-test with a 0.05 significance level is used for testing statistical significance of performance differences. Overall, the ensemble learning combining the fine-tuned RoBERTa and linguistic features achieves a significantly higher accuracy compared to all other models. These results provide an answer to our main research question regarding the existence of linguistic and grammatical differences between pseudo-science and pro-science articles.

• **Linguistic Features Importance**. We next perform an ablation study to evaluate the importance of the linguistic features and report the results in Table 3. To this end, we perform several model runs and each time remove a single set of features from the model and measure the impact on its performance. It can be observed that the Readability features contribute to a small improvement in the model performance. A more significant improvement is achieved using the lexical features. Finally, the POS tagging is the most important out of the linguistic features, as a result of differences in the writing style of pro-science and pseudo-science articles, with the latter using more emotive and loaded language that manifests in distinctive grammatical patterns.

6 CONCLUSION

In this work we proposed a novel large-scale dataset for identifying pseudo-science across a plethora of sources, which is substantially more diverse compared to previously available data. We presented a method combining both fine-tuned transformers and computational linguistics to distinguish between pseudo-science and pro-science content. We showed that our combined method achieves a significantly higher accuracy than each method in isolation, and demonstrated the effectiveness of grammatical tagging for identifying pseudo-science. As avenues for future work, we consider leveraging additional information, such as articles' images and video, in a multi-modal setting for improving model accuracy, as well as expanding the dataset into multiple languages for further applications.

REFERENCES

[1] Martin Curd; J A Cover. 1998. Philosophy of science : the central issues. W.W. Norton Co.

[2] Srijan Kumar and N. Shah. 2018. False Information on Web and Social Media: A Survey. *ArXiv* abs/1804.08559 (2018).

[3] Or Levi, Pedram Hosseini, Mona T. Diab, and David A. Broniatowski. 2019. Identifying Nuances in Fake News vs. Satire: Using Semantic and Linguistic Cues. *CoRR* abs/1910.01160 (2019). arXiv:1910.01160 http://arxiv.org/abs/1910.01160

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[5] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[6] Julio C. S. Reis, André Correia, Fabrício Murai, Adriano Veloso, and Fabrício Benevenuto. 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems* 34, 2 (2019), 76–81. https://doi.org/10.1109/MIS.2019.2899143

[7] Michal David Scott O. Lilienfeld, Rachel Ammirati. 2012. Distinguishing science from pseudoscience in school psychology: Science and scientific thinking as safeguards against human error. *Journal of School Psychology* (2012).

[8] Alexander Shvets. 2015. A Method of Automatic Detection of Pseudoscientific Publications. 533–539. http://dx.doi.org/10.1007/978-3-319-11310-4_46

[9] NawarathnaRuwan Nawarathna Thilina C. RajapakseRuwan. 2021. Pseudo-science Detection Using a Pre-trained Transformer Model with Intelligent ReLabeling. Proceedings of International Conference on Sustainable Expert Systems. http://dx.doi.org/10.1007/978-981-33-4355-9_25

[10] Alexandra Lefevre Rada Mihalcea Verónica Pérez-Rosas, Bennett Kleinberg. 2017. Automatic Detection of Fake News.

[11] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6