CS447 Literature Review: How do Artificial Intelligence and Natural Language Processing aid in Automatic Fake News Detection?

PRAVEEN KUMAR MURUGAIAH,
pkm4@illinois.edu

December 8, 2022

Abstract

In this literature review, we will focus on how various feature engineering techniques in text have been successfully utilized to aid Artificial Intelligence (AI) and Natural Language Processing (NLP) models to discriminate between fake and real news. We also systematically review five different papers on fake news detection in various domains and discuss the key modelling and evaluation methodologies used in great detail.

1 Introduction

Fake News poses serious and damaging consequences in a variety of fields ranging from healthcare and politics to national security by manipulating the thoughts and actions of people consuming them. The threat of fake news is growing rapidly with the increase in social media in recent years. While manual fact-checkers are often employed to check the truthfulness of the news in most online social media, the speed of fake news dissemination demands automatic fake news detection techniques as manual checking will be practically impossible. In this paper, we survey papers on how Machine learning (ML) and AI are applied to textual fake news to detect fake news by incorporating various features that not only improve the predictive capacity of the models but also provide a medium of understanding the working of the model. Moreover, we discuss the evaluation methods for these models in detail. We define fake news as any news that, intentionally or unintentionally, has the potential to deceive the readers consuming them. We specifically focus on political and healthcare related fake news, satire, hoax and propaganda which have the characteristics to confound users and also on neural fake news. While the former is real fake news found in news articles, existing open source news data sets and social media, the latter is machine-generated fake news which is constrained to match and many times exceed the expectations of real fake news in confounding the users. Our contributions are three-fold:

- We provide an overview of the topic and background information for easy understanding.
- We analyze the feature engineering techniques, models and evaluation methods used in the suggested papers.
- We categorize and summarize relevant findings from the proposed papers relevant to our research problem.

2 Topic and Motivation

Confidence in the mainstream media is decreasing all around the globe driving users to rely on alternative sources of information such as blogs and social media channels like Facebook, Twitter, Reddit etc. This has led to an increase in access to even more disguised, incomplete information. We as humans have inherent truth and confirmation bias with perceiving such information by either assuming all the news that we receive is trustworthy enough or just looking into what we want to see in the news excluding all other information. Such false news has more than seventy percentage probability of being shared online by users. Healthcare and politics are popular domains where such news spreads very rapidly and brings in more danger. Fake news related to Covid-19 and Ebola virus has had serious consequences in 2020 where people believed the veracity of news questioning the validity of vaccines preventing people from taking them. This could lead to the loss of lives. Similarly, in politics, such news can manipulate voters with false information incorporating negative sentiments. There have been a lot of efforts to identify and eliminate its spread in social media and other online platforms. Conventional systems require costly manual fact-checking based methods by human judges such as journalists but since viral fake news spread just within ten minutes of posting and hence such manual efforts are not at all effective. This demands the need for automatic fact-checking and fake news detection systems.

Unlike structured data where the data is organized in some tabular format with all relevant attributes at numerical, categorical levels, textual data is highly unstructured. Machine learning and AI modelling problems currently can only work on numerical and structured data feature Moreover, finding any insights in such unstructured data is very difficult requiring high technical expertise. News data is textual (unstructured) and in order to effectively mine the characteristics of the fake and real news, we need to dig deeper into the data and come up with a variety of attributes which could aid our analysis and interpretation for our machine learning models. Our research problem which we try to explain in this literature survey is to combine both these two problems into one by putting forth the question as to how can NLP based ML and AI systems help in the synthesis, extraction and analysis of features in detecting fake news in textual news data in an efficient way.

3 Background

In this section, we will briefly discuss the concepts and terminologies used for data preprocessing, modelling and evaluation that has been used in the papers surveyed.

3.1 Data Preprocessing

We mentioned earlier that machine learning-based models will work only on structured data representing numbers and textual data is unstructured. Input text data can be represented in a variety of ways which will be explained below.

3.1.1 Lingusitic Inquiry and Word Count (LIWC)

LIWC is defined as the percentage of the total words in the sentence that matches each of the words in categories of a predefined dictionary of words.

3.2 Machine Learning Models

There are a variety of supervised algorithms that could capture linear, non-linear space representations to correctly identify and separate fake and real news by the use of features.

3.2.1 Non-Neural: Linear regression

Linear regression models the linear relationship between the response variable (continuous) and one or more explanatory variables (continuous or discrete). The model learns the parameters by trying to fit a linear subspace that minimizes the squared errors (the error is the quantity left after subtracting the predicted and actual values).

3.2.2 Non-Neural: Ridge Regresser

Linear regression models usually suffer from multi-collinearity. i.e. the assumptions of predictive variables being independent of one another fail. Therefore, the model may overfit the data on training and the weights of features will no longer be acceptable for prediction. Hence, ridge regression uses an additional error term (for L2 regression) which punishes the model weights for being too large and hence, reduces the absolute feature weightage preventing overfitting.

3.2.3 Non-Neural: Maximum Entropy

The maximum entropy classifier is also a probabilistic classifier which works based on the principle of maximum entropy but it does not assume that the features are independent of each other, which makes it better than the Naive Bayes model.

3.2.4 Non-Neural: k-Nearest Neighbor Classifier

k-Nearest Neighbors (kNN) is a non-parametric supervised learning classification algorithm working on a clustering principle where the class prediction of a new point depends on the majority voting based on all the classes of some k-nearest points (distance usually measured in euclidean space) from that new point.

3.2.5 Non-Neural: Decision Tree and Random Forest Classifier

A decision tree is a flow chart of a hierarchical representation of the flow of decisions (if-else) from top to bottom. Each decision tree is constructed based on feature selection from the root node to the leaf node such that the impurity (or the class mix) in the leaf nodes (final prediction nodes) is minimized. While a decision tree uses all the attributes in the data for training, a random forest uses multiple small decision trees with just a subset of attributes and with randomly sampled data with replacement. Since a majority vote happens to decide on a class of the predicted point (aka. bagging), random forest classifiers are more powerful in predicting compared to a single decision tree.

3.2.6 Non-Neural: Support Vector Machines

Support vector Machines (SVM) work on the principle of maximum margin classification which tries to create a linear separation by projecting the data into a higher dimensional space with a high margin between the classes. Bidirectional Encoder Representations from Transformers (BERT)

is a transformer-based language model consisting of encoder layers and heads for self-attention pre-trained for language modelling and next-sentence prediction tasks. It learns contextual word embeddings while training.

3.2.7 Neural: GAN

Generative Adversarial Network (GAN) relies on two different and opposing machine learning models called the generator and discriminator. The generator tries to generate data using the statistics of the training dataset which could confuse the discriminator. The discriminator, on the other hand, discriminates real data and the data generated by the generator.

3.2.8 Neural: Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are regularized versions of fully connected MLPs which operate based on a shared weighted architecture with feature mappings and filters, thereby successfully detecting the local patterns in the input data very well.

4 Motivation of the Included Papers

As discussed earlier, we would like to identify fake news using natural language processing techniques by incorporating various features and also test them through evaluation.
Rashkin et al. (2017) states how news can deceive readers through stylistic lexicons and talks about satire, hoax and propaganda types of news. The reason to include this paper is that the papers talk about news types which are entirely not true or false and provide justifications for the usage of linguistic feature analysis to aid ML models to differentiate them better. This resonates with common sense that any news, even though it is true, will have some probability of containing false or deceiving information and we thought of including this paper to open readers' minds to this scenario.
Rubin et al. (2016) focuses on the satirical type of news which can hide the hidden truth in news employing humour or joke. This paper tries to find unique features in satirical news by incorporating five predictive features: absurdity, humor, grammar, negative effect and punctuation. The paper evaluates a variety of these feature combinations and discusses how each combination helps the model distinguish satire from non-satire news. The readers could gain insights on cues to detect satire which indirectly confuses readers in the real meaning or truth buried in the statements.
Aich et al. (2022) conducts a feature-based study with 3 misinformation feature categories divided into 21 unique features. The goal of the paper is to effectively utilize these features to interpret how these linguistic features can be exploited by neural fake news generators at varied levels of Generative Adversarial Network's complexity. Since nowadays, a lot of the fake news is generated by such fake news generators constrained with some user-defined limitations to make it sound too real compared to true fake news, readers reading this paper could be able to gain a comprehensive understanding as to how fake news from such generators can be detected with confidence and how the features described could aid in the successful detection.
Oshikawa et al. (2020) provides a brief survey on different other papers on AI for fake news detection using NLP and summarizes a variety of datasets used in these papers, methods such as data pre-processing, models including both non-neural machine learning models and neural models such as recurrent neural networks and attention based mechanism used and discusses the requirements of any fake news detection system. The paper also provides recommendations and suggestions as to what new methods and approaches can be tried out in the future for fake news detection using NLP

techniques. The readers can gain an overall understanding of what is fake news and how it is different from real news, what existing data sources are used in research, what data modelling pipelines are used in building such models, which would serve as a source of inspiration and introductory knowledge on NLP research for fake news detection.
Hansen et al. (2021) is very different from the above-mentioned papers. The papers mentioned so far describe more about the features and modelling techniques and little on evaluation but this paper contributes to evaluating news (claims) based on evidence (for that particular claim) supporting or opposing them. It provides readers with an answer to the question "Given a claim and a set of evidence, are machine learning models trying to learn useful information (reasoning) from the evidence?", thereby providing a means of understanding the evaluation of claim and evidence combinations.

5 Summarization/Description of the Discussed Papers

Oshikawa et al. (2020) discusses papers by Rashkin et al. (2017) that consider 2 different datasets used as benchmarks for any fake news detection tasks recently. LIAR is a dataset from the Politifact website containing news related to politics with speakers' statements and speakers' details. It consists of 12,836 short claim statements each with a 6-grade level of truthfulness and with information related to the content of the claim, the speaker of the claim and the party affiliated. FEVER is a dataset used for fact-checking. It consists of 185,445 claims from Wikipedia tagged with three labels denoting whether that particular claim is being supported, opposed or with not enough information to support or oppose. Rashkin et al. (2017) proposes 2 different modelling strategies. One with a single LSTM model with simple word embedding and another with two LSTM models with a simple Word2Vec embedding to one side of the model and LIWC feature vectors to the other side of the model. Wang (2017) proposes a Convolutional Neural Network model (CNN) where the input features were max pooled textual representations from a bi-directional LSTM model. This combination was used to extract features from various metadata varieties including graphs. Karimi et al. (2018) also proposes a CNN + LSTM-based model where the CNN analyzes local data patterns and features in the claims and LSTM identifies any temporal dependencies in the input data. Finally, the output hidden states are concatenated with a feed-forward neural network layer. Long et al. (2017) proposes attention-based mechanisms including the speaker's name and the statements of the speaker as raw features as input to LSTM again.

Rashkin et al. (2017) considers hoax, satire and propaganda as fake news, as mentioned earlier, as the truth in the news is not always apparent to people with bias. This paper focuses on the linguistic attributes used in various political news and social media. The paper defines a hoax as a paranoia-fueled story, satire as a comedy mimicking reality but with hints to make the joke apparent and propaganda as misleading news that makes people believe some political agenda. Four different lexicons were proposed as features to differentiate satire, hoax and propaganda. The first lexicon is the LIWC which is predominantly used in social science research. The second lexicon is the sentiment lexicon with strong and weak subjective words to give the news a more dramatic effect. The third lexicon is the hedging lexicon which uses obscure language and the final lexicon is the intensifying lexicon from wiktionary used to attract news readers. These lexicons are broken down into individual features such as first-person singular, second-person (You), Swear words, Modal Adverb, Action Adverb, Sexual, See, Negation, Strong subjective, Hedge, Superlatives, Number, Hear, Money, Assertive and Comparatives. The ratio of the presence of these features in both fake

news and real news is being compared through machine learning modelling. Different models with various input feature representations were tried out. First was a simple LSTM model which is compared with Maximum Entropy and Naive Bayes baseline models with 2 different inputs - one with tf-idf and another with LIWC and tf-idf combined. Another was again an LSTM model with word embeddings from Glove as input where the output of the LSTM model is concatenated with LIWC feature vectors. The embeddings are 100 dimensional fine-tuned in the training phase and the model consists of a 300-dimensional hidden state with a batch size of 64, optimizer as ADAM and number of epochs as 100. The objective of the model is to minimize the categorical cross-entropy loss. The final class prediction is on 6 point scale - More true (true, mostly true and half true) and More false (mostly false, false and pants on fire) which could also be combined for just 2 classes (more true and more false). The models were evaluated by using a variety of mix of news sources as a training set and a completely different news source for testing with all the quotes of the same speakers in a single set.

The paper Rubin et al. (2016) conducts a feature-based study on 21 features on GANs to find the feature stability with increasing levels of model vulnerability on health data from the Buzzfeed news website. The features are stylistic (quotes, punctuation, unique punctuation types, exclamatory marks, stop words, camel case, negation, proper nouns, user mentions, hashtags, misspelt words, out-of-vocabulary words, nouns, past tense, verbs and interrogative words), complexity (word count, mean word length, Type-Token-Ratio (TTR) and MTLD) and psychological (sentiment score). Six classical machine learning models (Linear regression, SVM, ridge regression, KNN, Decision tree and Random Forest) on balanced toy data with 200 samples from the "Telling a lie" dataset containing an equal number of true and fake news. The best model out of these was selected as a baseline model and the performance is compared with a GAN model. The Generator in GAN is a 2-layered LSTM with binary cross entropy as a loss function and auto-regressive language generation as an objective task. The constraints are made such that the model tries to generate fake news from BuzzFeed's real news data. The number of epochs is assessed for feature vulnerability. The discriminator is a 3-layered CNN with leaky ReLU as an activation function having a sigmoid layer in the final layer to classify real and fake news. Input for this discriminator is from the final hidden layer of the generator with all the twenty one unique features selected above. The GAN evaluation is a 2 step process. Firstly, KNN with the 21 features is experimented on (a) 30 random news articles from the Buzzfeed real news data and 30 random news articles from the Buzzfeed fake news data and (b) 30 random news articles from the Buzzfeed real news data and 30 articles from GAN at some optimal epoch level. Secondly, GAN at 10, 20 and 30 epochs are tested for 30 random news articles from the Buzzfeed real news data and 30 articles of GAN to make the model challenged more to the increasing levels of difficulty to detect fake news.

Aich et al. (2022) argues that satire is hard to pin down in news and many works of literature. It mentions the usage of word-level features such as headlines, profanity and slang and SVMs on Bag-of-Words with feature weighting methods and also the corpus level relative features such as cosine similarity and tf-idf, individual words and n-grams as features for models. Here, the baseline model was chosen as the model with tf-idf on bigrams with additional features from semantics. The paper proposes and tests five different satirical features such as absurdity, humour, grammar, negative effect and punctuation. Absurdity (abs) was detected by the sudden introduction of newly named entities in the last sentence of the article. POS tagging was implemented to identify this. Humour (hum) was detected based on the presence of opposing scripts identifies using punch line

detection based on the word-word similarity score. Grammar (gram) was detected by comparing LIWC 2015 dictionaries which have all POS tagging and grammar with the set of normalized term frequencies in the given news article. Finally, negative effects (neg) were detected through the presence of negative words and punctuation through the presence and count of unique punctuation. All these features are fed as input to the SVM model with 10-fold cross-validation with a binary classification objective. All the features are normalized by sentence length before input after the elimination of stop words.

Hansen et al. (2021) focuses just on how claim and evidence would be helpful for models to learn and asks a question of whether a model with both claim and evidence combined is better than either of a model with just claim or a model with just the evidence. The dataset consists of two different political fact-check data. Features such as hand-crafted features, BOW representations and deep learning based contextual and non-contextual embedding were used on three different models. Tf-idf weighted BOW-based features on random forest classifier with claims and evidence was concatenated together is compared with scenarios where just only claim was used and only evidence was used. The model used a variety of parameters for tuning such as the number of trees, minimum number of samples per split and number of samples in the leaf node. Glove-based LSTM with a pre-trained claim and evidence embedding using attention-based bidirectional LSTM encoders and having cross-entropy loss. Similar to the previous one, comparison again happens with only claims as input embedded features and only evidence as input embedded features. This model tries various parameter combinations such as learning rate, batch size, number of hidden layers to be used and the number of hidden dimensions. A BERT-based model with token encoding for claim and evidence text feature representations with cross-entropy loss was compared with the model with just claim and evidence alone. This model tries out a variety of learning rate and batch size parameters. Claims are taken from two different political fact-checking datasets (Politifact and Snopes) and evidence is collected from google searched through google search API where the top 10 results are collected as evidences.

6 Assessment, Interpretation, Analysis, or Discussion of the Individual Papers

In (Oshikawa et al., 2020), it is found that the LSTM model performs very well on LIAR and FEVER datasets with high accuracy. The dual LSTM with Word2Vec embeddings and LIWC feature vectors has even higher accuracy than the Naive Bayes classifier and maximum entropy models. Attention-based models perform better than all the above models increasing accuracy by 3 percentage. The results show that LSTM achieves high accuracy than CNN and additional metadata features had been very important.

Inferences and Suggestions: The paper discusses binary classification-based methods but we also need to explore models with 6-way classifications such as true, mostly true and half true, mostly false, false and pants on fire. Also, past literature surveyed in this paper assumes some publishers as always authentic but that assumption may not always hold true. Hence, new methods can be put forward where the bias with particular authors or publishers is eliminated. The paper also doubts if hand-crafted features, which work very well in classical machine learning models, will also work with neural networks. We will see upcoming papers where this doubt is clarified. We could also use metadata and content-based approaches to improve the performance of models while also reducing

noise.

In Rashkin et al. (2017), the LSTM model with just Word2Vec as input features outperforms all other models. LIWC as features improves performance in the multinomial Naive Bayes and Maximum entropy models whereas in neural networks such as LSTM, LIWC does not improve performance at all and such new lexical information from LIWC features had been redundant. This implies that LIWC works very well in classical machine learning model in terms of aiding the models with more information to discriminate between fake news and real news in varying levels such as More true (true, mostly true and half true) and More false (mostly false, false and pants on fire) while in neural networks the model learns noisy information more than the useful lexical information from LIWC features. The evaluation of the 21 features was done using a statistical significance test with Welsh t-test and the ratio of these features' prominence in both fake and real news was obtained. The analysis suggests that 1st and 2nd person pronouns are more prominent in fake news than real. This could be because news writers try to remain indifferent to news and remove the usage of personal language in their articles. Similarly, subjectives, superlatives and modal adverbs are more dominant in fake news, which could be due to the characteristic of fake news having words to describe and make things appear even bigger and have a greater impact than reality. Moreover, concrete figure words such as comparatives, money and numbers are more dominant in true news implying that solid real facts are always present in real news than fake news. Trusted news sources have more assertive words and are less vague than fake news. Satire has more adverb usage, hoax has fewer superlatives and comparatives and propaganda has relatively high assertive verbs and superlatives. The results also show that lexical features such as swear, 2nd person, 1st person singular, modal adverb, action adverb, manner adverb, sexual, see, negation, strong subjective, hedge, superlatives and weak subjective have a ratio of being present in fake news than real news is more than one while other lexical features such as numbers, hear, money, assertive, comparatives have the ration less than one.

Suggestions and Future Work: Since we just focus on the political domain in this paper, we need to explore how these input features and lexical features would work for other domains such as healthcare, etc.

Rubin et al. (2016) analyzes tweets and news articles for misinformation features (stylistic, complexity and psychological). It was found that tweets with fake news contain less vocabulary, longer sentence length and double user mentions than true news with 62 percentage more exclamation marks. In articles, fake news articles are shorter with longer titles. Fake news articles also have more capitalized words, proper nouns, verbs, past tense words, personal pronouns, self-referential terms and adverbs and fewer nouns, stop words and vocabulary with less punctuation. KNN model is selected as a baseline with an accuracy of 97 percent and an f1 score of 0.9. In GANs, there is a significant performance drop when training happens for more epochs and the model gets challenging information. The results indicate that removing comparative features did not affect the model's performance as the misinformation grows more challenging. This implies that the comparative features add noise more and more as the training happens more and more. At 30 epochs, models without comparative features show higher performance than with those features. But the removal of stylistic features makes the strongest performance drop at all epochs (10, 20 and 30). This indicates that stylistic features such as quotes, punctuation, proper nouns, noun, grammar, etc. help a lot in differentiating fake and real news. Removal of psychological features makes a performance drop but it is not very clear why that happens and needs further investigation which is not covered in this paper.

Inferences and Future Work: It is found that both correctly predicted fake news and wrongly pre-

dicted fake news have polar terms indicating that there is a scope of research for analyzing stances, opinions and hate speech in the news articles. Moreover, correctly identified misinformation supposedly had a higher frequency of punctuation and more usage of numbers. From this, we can infer that we could, in the future, select features indicating the frequency of numbers or digits. Additional research can be conducted following this paper such as how multi-GANs may work (rather than single GAN), how domains other than politics and health may affect the model and how other language articles (apart from English) could influence models and features.

In Aich et al. (2022), preliminary analysis of features suggests that headlines are very helpful in detecting satire and also the final punch line of the article which often contains absurdity in story is deemed to be highly important. Satirical news also contains high-frequency usage of slang, swear words and longer sentence length accounting for more punctuation marks. The best model was found to be SVM with Absurdity, Grammar and Punctuation with an an 84 percent recall rate, 90 percent precision and 87 percent f1 score. Profanity and slang did not contribute to the model performance. Similarly, maximum precision is obtained when Grammar and punctuation are used and maximum recall is obtained when just absurdity is used.

Inferences: We can infer that POS tags can very well help in satire detection implying that the rhetorical component of satire aids well. The additional inference is that absurdity helps our model which is detected by considering the presence of non co-occurring entities in the last sentence of the article. Also, humour is identified by the presence of latent terms in the article with shifting reference frames. This means that latent features help in detecting satire as they are used by authors to separate the joke and narration content in such satire articles. We can also infer that positive semantic orientation does not help the model in anyways as much as negative semantic orientation helps to detect satire.

In Hansen et al. (2021), the models are evaluated by training on the Snopes dataset and tested on the Politifact dataset to check the robustness of the model. The best performance is obtained when the claim is entirely ignored from analysis and only evidence is used. This implies that there is a strong predictive signal in evidence alone. The best model is the BERT trained only on evidence.

Inferences and Suggestions: From the results, we can infer that evidence in itself has a predictive capacity. Models with both claims and evidence just exploit the signals from evidence. Additionally, we can see a flaw in the evidence-collection approach. The suggested method using google's top 10 evidence for claims will not be a desirable approach since searching for two different opposing claims may show the same set or subsets of evidence. The results also show that performance gradually decreases while removing evidence in both top-down (removing most relevant evidence first or removing top 3 google search results) and bottom-up (vice versa) in both Random forest and LSTM. BERT shows a very small performance decrease when the claim is removed implying claims are not useful in predictions. Additionally, we can also analyze how the presence of metadata features in both claim and evidence would affect the model.

7 Contribution of each paper to the topic/research question

In addressing the research question "How does NLP and AI aid in fake news detection?", the paper Oshikawa et al. (2020) discusses a variety of data pre-processing, input feature representation and modelling strategies that have been implemented in previous research and suggests the use of handcrafted features in settings of both non deep learning and deep learning modelling. This serves as a starting point or precursor to our next research as this paper forms a foundation for understanding the basic outline research that happened in NLP on fake news identification tasks.

Rashkin et al. (2017) shows how various lexical features in NLP contribute to identifying satire, hoax and propaganda news which are considered fake/deceiving news in this survey.
Rubin et al. (2016) analyzes how misinformation features such as stylistic features, complexity features and psychological features aid in detecting fake news both real fake news and generated fake news from GAN at varying levels of difficulty.
Aich et al. (2022) portrays how predictive features such as absurdity, humour, grammar, negative effect and punctuation could aid in detecting satirical cues in news articles.
Hansen et al. (2021) is very different as it comes up with a question on whether the evidence for fake news detection tasks in isolation helps in the model's predictive capacity which goes a little off from our research question but also provides a modelling approach to address this question using claim and evidence represented as input features, claim alone represented as input features and evidence alone represented as input features in 3 different models.

8 Overall discussion and conclusion

Oshikawa et al. (2020) suggests varied levels of true/false and puts forth the research question on how handcrafted features would be helpful in deep learning models to differentiate fake and real news. With this question in mind, we look for Rashkin et al. (2017), which uses fact-checking on a 6-point scale (true, mostly true, half true, mostly false, false and pants on fire) and proves that handcrafted features are not performing well in neural models like LSTM but performs better on non-neural models like Naive Bayes and maximum entropy classifier in detecting satire, hoax and propaganda. Rubin et al. (2016) talks about neural fake news as they are more common nowadays with the increasing use of social media and provides key features and models which could help in discriminating both real fake news and generated fake news from real news. The latter papers questions the validity of features and models on other domains apart from politics and health and this question was analyzed in Aich et al. (2022) where training happens with data from one domain and testing happens with data from a completely different domain. It also considers key features such as absurdity, humour, grammar, negative effect and punctuation in improving the performance of satire news detection. Finally, we would like to evaluate whether claims in fact-checking tasks (similar to fake news detection) could effectively help in identifying true facts. The previous four papers consider labelled true or fake news but do not contain any claims and in this paper Hansen et al. (2021), it is found that evidence alone could prove to be a strong predictor and having claims is not preferred which reinforces the methodologies and data we used in the previous four papers.

References

Aich, A., Bhattacharya, S., and Parde, N. (2022). Demystifying neural fake news via linguistic feature-based interpretation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6586–6599, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hansen, C., Hansen, C., and Chaves Lima, L. (2021). Automatic fake news detection: Are models learning to reason? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 80–86, Online. Association for Computational Linguistics.

Karimi, H., Roy, P., Saba-Sadiya, S., and Tang, J. (2018). Multi-source multi-class fake news detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1546–1557, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Long, Y., Lu, Q., Xiang, R., Li, M., and Huang, C.-R. (2017). Fake news detection through multi-perspective speaker profiles. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 252–256, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Oshikawa, R., Qian, J., and Wang, W. Y. (2020). A survey on natural language processing for fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (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.

Rubin, V., Conroy, N., Chen, Y., and Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Wang, W. Y. (2017). "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.