Semantic Abstraction: On the Generalization of Fake News Detection

A0216121Y, A0217701N, A0218441L, A0230337N, A0235341N, A0235393B
Group 02
Mentored by Xian He
{e0538222, e0543737, e0544477, e0694459, e0727341, e0727393}@u.nus.edu

Abstract

Fake news presents a significant detriment to modern society, deceiving various consumers. Fake news classification is thus a major concern to curb the consequences of the proliferation of fake news. Prior work investigates fake news classification with different models, as well as the ability of such models to generalize on unseen data. In this paper, we propose two novel generalization techniques, synonym augmentation and NER masking, to help with model generalization performance. We explore the effects of these proposed methods on different model types (LR, CNN, Bi-LSTM, DistilBERT). Our experiments benchmark performance against prior work, and show promising results, with some models showing improved generalization performance on an external dataset.

1 Introduction

The proliferation of fake news represents a growing threat to society, posing significant challenges to individuals and communities around the world. Statistics have shown that 70% of consumers perceive fake news as a significant issue in contemporary society [1]. Though work has been done to optimize models to detect fake news, further advancements and methods could be made to improve the detection of fake news on unseen data.

We propose the following means of generalization in improving the fake news detection on external datasets:

- Synonym-based augmentation of corpora.
- Named Entity Recognition (NER) masking.

To our knowledge, synonym augmentation and NER masking in fake news detection has not been researched in literature. We make the following contributions:

- We explore how the proposed synonym augmentation of corpora affects the generalization of models in fake news detection
- We explore the relation between named entities and fake news detection
- We analyse how the proposed methods affect the performance of different models in classifying fake news as satire, hoax, propaganda, and trusted

2 Related Work

In exploring the generalization of fake news detection over different contexts, we examine prior work proposing various models for fake news detection. Additionally, we review prior work analyzing the extraction and use of different features in improving performance on fake news classification tasks.

2.1 n-gram based models

Ahmed et al., 2017 explores the use of n-gram models to detect fake review and news content. The authors propose a machine learning classification model that utilizes text analysis of n-gram features and term frequency metrics, *tf-idf* and *tf*. They used word-based n-grams to capture context and generate features with which to classify documents, generating n-gram frequency profiles to differentiate fake content. They also examine the effects of n-gram lengths ranging from $n = 1$ to $n = 4$ and the number of selected top features on the performance of different classification algorithms using several baseline word-based n-gram features.

Analysis of the datasets used show that fake content contained more function and content words, while truthful content contained more nouns and adjectives. In particular, fake news contained more adverbs and verbs, and in one dataset used less verbs and adjectives.

Their results show that linear-based classifiers like linear SVM, SDG and LR achieve better results than non-linear classifiers, with linear SVM

[1] https://www2.deloitte.com/xe/en/insights/industry/technology/study-shows-news-consumers-consider-fake-news-a-big-problem.html

achieving a 92% accuracy, though non-linear classifiers like DT achieved a respectable 89% accuracy. Longer n-grams (trigram, 4-gram) serve to lower accuracy, with unigrams and bigrams performing well, though trigrams perform better on fake news compared to fake reviews. As a feature extraction method, *tf-idf* also outperformed *tf*. Increased feature sizes at 10000, 50000 features, also achieved higher accuracy.

2.2 CNN, Bi-LSTM based models

Roy et al., 2018 develops a deep multi-label classifier using CNN and Bi-LSTM based representations that are fed into an MLP model for classification of fake news from the LIAR dataset into six classes of truthfulness.

The authors hypothesize that the ability of CNN to efficiently capture hidden features would be able to detect hidden features and information to judge authenticity. They suggest that these hidden features, in conjunction with the sequential information in statements captured by Bi-LSTM, would be necessary in classifying authenticity.

Relations between various attributes were identified to contribute towards the labeling of statements, such as between statement, statement type and context. Each of these relations were fed into individual layers for the deep networks to comprehend and extract. The proposed model passes each attribute through different layers instead of through separate LSTMs, and attains an overall accuracy of 44.87%, outperforming existing SOTA methods.

2.3 Alternative features

Wang et al., 2018 explores feature extraction from news and social media, and analyses the use of these features in classifying articles. Based on assumptions on potential indicators of truthfulness, the authors identified the following feature categories: named entities, headline, sentiment lexicon and subjectivity lexicon.

Evaluating on the SHPT (Rashkin et al., 2017) and PolitiFact datasets, they found that using all tokens as a feature tends to outperform the proposed alternative features, though proposed features are able to achieve reasonable performance at reduced feature dimensions.

2.4 Readability as a feature

Santos et al., 2020 studies the significance of readability in detecting fake news in Brazilian Portuguese. On top of previous approaches involving network information, linguistic features and lexical measures, the authors propose readability analysis as a means of extracting features to improve fake news classification tasks.

Readability features explored were categorised under *Classic*, *Cohesion* and *Psycholinguistic* features. Results show that such features alone were able to achieve up to 92% accuracy in fake news classification, though using them in conjunction with other features yield better results.

3 Methodology

3.1 Data

Three separate corpora were used for training and testing the models. All corpora used are multi-class datasets that classify news with the SHPT taxonomy (*Satire*, *Hoax*, *Propaganda*, *Trusted*) introduced in Rashkin et al., 2017.

3.1.1 Train corpus

The train corpus was sourced from an ACL 2021 paper (Hu et al., 2021) which classifies news into 4 different classes: satire (14047), hoax (6942), propaganda (17870) and trusted news (9995). An 80:20 split of the train corpus forms the training set and validation set.

3.1.2 Test corpus 1

Test corpus 1 was sourced from the same ACL 2021 paper (Hu et al., 2021) and uses a separate set of news samples taken from similar sources as a balanced test set. There are 750 test samples each for satire, hoax, propaganda and trusted news.

3.1.3 Test corpus 2

The samples in test corpus 2 are taken from several different sources, OpenSources for satire and trusted,[2] HackNews Datathon for propaganda,[3] and the LIAR dataset for hoax samples.[4] 3500 samples of each label were randomly chosen and combined to form a test dataset.[5] This dataset is primarily used for cross-dataset validation on external datasets.

[2] https://huggingface.co/datasets/andyP/fake_news_en_opensources

[3] https://github.com/leereak/propaganda-detection

[4] https://huggingface.co/datasets/liar

[5] https://github.com/autumn-sonata/CS4248-Project/tree/main/synonym-creation

Label	Count	Source
Satire	3500	OpenSources
Hoax	3500	LIAR
Propaganda	3500	HackNews Datathon
Reliable	3500	OpenSources

3.2 Evaluation metrics

The evaluation of each model is primarily based on the F1-score. An overall macro F1-score was calculated, along with F1-scores gathered for individual labels. F1-score was chosen to evaluate binary classification performance (e.g. satire or not satire) to measure how well each model performs in classifying each label. The overall F1-score was also used to benchmark each model against the performance of models proposed in prior works. Precision and recall metrics were also gathered for further evaluation of model performance and effectiveness of the proposed generalization methods.

3.3 Proposed Generalization Methods

We propose the following as a means of generalizing the detection and classification of fake news.

3.3.1 Synonym Augmentation

Synonyms are words or expressions that have a similar semantic meaning. We propose the augmentation of training corpora with synonyms as a means of improving generalization.

Synonym augmentation of training corpora was done by removing punctuation from examples in the corpora, followed by the duplication of a randomly selected examples, replacing up to 10 verbs or adverbs with synonyms based on the *word2vec-google-news-300* pretrained Word2Vec model with a lower bounded similarity threshold of 0.65. This was done until each of the number of examples for each label reaches 110% of the initial number of examples of the majority class. A new synonym-augmented train corpus is thus created,[6] with the following statistics on the counts of each label:

Label	Before Count	After Count
Satire	14047	19657
Hoax	6942	19657
Propaganda	17870	19657
Reliable	9995	19657

[6] https://github.com/autumn-sonata/CS4248-Project/tree/main/NER-masking

3.3.2 NER Masking

NER masking is applied during preprocessing. We postulate that named entities contribute to the contextualization of news sources and skew the classification of a text based on those named entities. For example, an entity like Donald Trump might skew the classification of an example as a hoax. NER masking masks all tokens deemed as named entities as tokens of predefined categories (e.g. Donald Trump → <PERSON>) prior to being passed downstream for feature engineering and modelling.

3.4 Preprocessing

Preprocessing is done only for the Logistic Regression model. The order of the preprocessing steps are as follows: Tokenization → Punctuation removal → Stopword removal → Lowercasing.

Certain preprocessing methods such as lemmatization were omitted as they might obscure provocative wording that might help classify fake news.

Punctuation removal and lowercasing enables a smaller dimensional space when looking at similar words in the word embedding space. Stopword removal helps prevent features like *tf-idf* from potentially prioritizing stopwords as important words (e.g., in the case that some stopwords appear frequently in only a few documents), especially in long text examples in the train corpus.

4 Experiments

A Python 3.11.8 environment was used for all experiments. Where applicable, model training and development was done on Kaggle with P100 GPUs, and used CUDA 10.8. Precision, Recall and F1-score metrics were collected to evaluate performance.

4.1 Setup

Fine-tuning of DistilBERT on the classification task was done with a batch size of 32, and the AdamW optimizer with a learning rate of 2e-5, over 5 epochs.

The train corpus is split into (80%) training and (20%) validation sets. Feature extraction is done based on model type, and each model is trained on the training set. Each model is trained on the training set, and metrics for prediction on the 20% validation set and test corpus 1 are collected for evaluation of performance on the standard multi-class classification task.

Each model is then trained on the full train corpus, and metrics for prediction on test corpus 1 and 2 are collected for cross-dataset validation. These metrics are used to evaluate the ability of each model to generalize on unseen data in external datasets. The proposed measures for generalization of multi-class classification of URLs are then applied, before training each model again and predicting on test corpus 1 and 2, to evaluate the effectiveness of the proposed measures

4.2 Models

The following models were explored:

4.2.1 Logistic Regression [LR]

A Logistic Regression model was chosen due to its performance when compared to other models such as RandomForestClassifier and MultinomialNaiveBayes in a preliminary test using *tf-idf* as the sole feature with no preprocessing on the LUN dataset.

The proposed LR model uses features that are predominantly motivated by prior work on fake news detection. *tf-idf* and *tf* on unigrams and bigrams was motivated by Ahmed et al. (2017). Sentiment analysis was cited as a predominant feature in Wang et al. (2018). Verb counting counts all available POS tags for verbs. Feature extraction is also motivated by the different classification labels. Propaganda tends to use emotional appeal to elicit a certain response from the reader, so sentimental analysis is used as a feature. Carrasco-Farré, 2022 finds misinformation is easier to read and less lexically diverse than trusted news motivating the use of Flesch Kincaid to measure readability of news articles.

Additionally, we utilize Word2Vec embeddings and sum the top 5 words by *tf-idf* value in order to approximate the document in a high dimensional vector space.

Other features like exclamation, question mark count and the number of sentences were also explored. We stuck to the initial 6 features as listed above due to the better classification on specific labels. Feature Selection was done using *SelectPercentile* algorithm with top 10% of features chosen from the feature set. Scoring was done on each feature based on the ANOVA F-test between labels and features.

4.2.2 Convolutional Neural Network [CNN]

A Convolutional Neural Network was used as a convolutional filter of stride n on the embedding matrix tended to act similarly to how an n-gram acts, capturing the context and semantic information from a sentence. These however are much more computationally faster than using n-grams and as a result we can generate and train such filters for the model at low compute cost, with each looking for a particular semantic feature. The usage of max pooling also reduces the data involved as we can ignore the results that point to a lack of presence of the feature we are looking for and just consider the max result as the net probability of feature presence. This helps distill a large number of features from the corpus into a relatively compact form. The choice of a CNN is backed by the results of Kim (2014) which showed the effectiveness of CNNs in sentence classification. However, pre-trained word embedding vectors were not used in favor of training the word embeddings to get better embeddings that are more appropriate to the context of fake news classification.

4.2.3 Bidirectional LSTM [BiLSTM]

A Bidirectional Long Short-Term Memory (BiLSTM) model was used, and GloVe embeddings was employed for word embeddings. BiLSTM models excel at capturing contextual information and patterns throughout the entire sequence. This helps it to understand long term dependencies over extended sequences, which is suitable for this task where inputs are long sequences with median run length of 274. The bidirectional nature of the model helps it to capture context from both past and future sub-sequences in the sequence, making it suitable for tasks like fake news classification where the input is finite in length.

GloVe embeddings were chosen over training custom embeddings as they are marginally more effective while saving considerable training time and resources compared to training embeddings from scratch. A relatively high dimensionality of 300 was chosen for the GloVe embedding to better capturing semantic nuances and contextual information to make up for the trade-off of losing the specificity of an embedding trained specifically for the context of fake news classification.

This choice of BiLSTm with GloVe was motivated by Wang (2023), who found that the bidirectional nature of the model made its performance marginally better than models of similar complexity such as TextCNN.

4.2.4 DistilBERT

DistilBERT is a distilled version of the larger BERT model, pretrained on the same corpus, with the aim of learning a similar representation while being faster at downstream tasks. The uncased version was used with the aim of ignoring casing and focusing on the semantic features of the news, along with *DistilBertForSequenceClassification*,[7] which adds a classification head on top of the model for use in classification tasks. The choice of this model is in part motivated by the results in Vaibhav et al., 2019, which leverages BERT to extract sentence level representations in news documents. The aim in using DistilBERT is to explore performance with the distilled BERT representation, as well as analyse any potential deterioration in the learned representation caused by the pretraining objectives of DistilBERT, specifically, the objective to return the same probabilities as the BERT model. An additional goal is to explore the extent to which the classification of fake news is able to leverage transfer learning from the representation pretrained on BookCorpus.[8]

4.3 LUN Results

The results on the train corpus and test corpus 1 are discussed in this section. Table 1 shows the F1-scores of models proposed in Rashkin et al., 2017 and Vaibhav et al., 2019, as well as the F1-scores of our models. Model training and evaluation follow the methodology of the referenced works, with the 80:20 split of the train corpus, and the models being trained on the 80% training set, and evaluated on the 20% validation set (**LUN-dev**), and test corpus 1 (**LUN-test**).

The performance of our proposed models largely align with thos of Vaibhav et al., 2019. Our CNN model exhibit only a marginal difference of approximately 1% in F1-score. DistilBERT, despite being lightweight, demonstrates comparable performance to the BERT-based as expected. The BiLSTM model surpasses the LSTM model, which can be attributed to the additional reversed LSTM, enabling the model to consider future context. This observation aligns with Wang, 2023, although the performance disparity is less pronounced in their study.

[7] https://huggingface.co/docs/transformers/en/model_doc/distilbert#transformers.DistilBertForSequenceClassification

[8] https://huggingface.co/distilbert/distilbert-base-uncased

Model	LUN-dev	LUN-test
(Rashkin et al., 2017)	91.0	65.0
(Vaibhav et al., 2019)		
CNN	96.48	54.04
LSTM	88.75	55.05
BERT	95.07	54.87
GCN	96.76	65.0
GCN + Attn	97.57	67.08
GAT	97.28	65.51
GAT + 2 Attn Heads	97.82	66.95
Our Models		
LR	95.93	71.58
CNN	96.63	54.29
Bi-LSTM	95.11	73.76
DistilBERT	99.45	54.10

Table 1: LUN 4-way classification F1-scores

4.4 Cross-Dataset Validation Results

The cross-dataset validation results on test corpus 2 is discussed in this section. Models are trained on the entirety of the train corpus, and evaluated on test corpus 2 for cross-dataset validation. Table 3 shows the performance of each model without any generalization methods applied, Table 4 shows model performance with synonym augmentation, and Table 5 shows model performance with NER masking.

5 Discussion

5.1 Test Corpus Limitations

When we compare Table 2 to Table 3, we see that the F1-score in classifying every label across all models drops when tested on the external dataset compared to the LUN test set. This could be attributed to the labelling in the external dataset. The unseen test corpus (test corpus 2) is created from multiple sources to increase diversity for cross-dataset validation. However, labels from the sources were only correlated with the 4 labels in our training set as per Roy et al. (2018), and not directly labelled with those 4 labels. Thus, "Half-true" and "Barely-true" labels were mapped as propaganda articles, and "False" and "Pants-fire" were mapped as hoax articles as per the the paper. However, satire and trusted news articles were taken from OpenSources, which was already vetted and cleaned on HuggingFace.

This could be one of the factors that led to the

Model	Class	Precision	Recall	F1-score
LR	*	74.10	74.16	73.93
	S	71.89	83.98	77.46
	H	82.71	47.20	60.10
	P	62.77	74.40	68.09
	T	79.18	85.20	82.08
CNN	*	68.50	61.97	59.55
	S	91.12	52.0	66.21
	H	59.85	21.07	31.16
	P	46.54	94.13	62.28
	T	76.49	80.67	78.52
Bi-LSTM	*	75.28	74.46	73.77
	S	82.19	77.03	79.53
	H	74.85	48.40	58.79
	P	62.97	81.87	71.19
	T	81.12	90.53	85.57
DistilBERT	*	54.55	58.49	53.17
	S	90.06	58.08	70.62
	H	68.28	75.47	71.69
	P	10.19	2.93	4.55
	T	49.69	97.47	65.83

Table 2: LUN-test 4-way classification scores (full)

Model	Class	Precision	Recall	F1-score
LR	*	43.83	48.77	44.82
	S	57.57	61.47	59.46
	H	16.17	4.63	7.20
	P	40.02	63.03	48.96
	T	61.56	65.94	63.68
CNN	*	40.07	38.43	35.85
	S	46.59	25.74	33.16
	H	28.14	10.0	14.76
	P	30.57	65.14	41.61
	T	54.98	52.83	53.88
Bi-LSTM	*	45.80	47.11	45.53
	S	50.18	47.96	49.04
	H	32.35	17.17	22.43
	P	36.88	50.69	42.70
	T	63.78	72.66	67.93
DistilBERT	*	45.25	38.54	33.83
	S	66.12	29.75	41.04
	H	30.75	10.58	15.74
	P	50.88	20.83	29.56
	T	33.25	93.0	48.99

Table 3: Cross dataset 4-way classification scores

evaluated scores for each label in Table 3. On average, all models performed worse when classifying hoax and propaganda compared to satire and trusted labelled news, which could be attributed to inaccurate mapping in the test corpus. This was especially the case for the hoax label, where the highest scoring model on the external dataset was the Bi-LSTM with a 22.43 F1-score. A self-labelled dataset could alleviate this problem.

5.2 Classification of Hoax news articles

The dip in performance for classifying hoax articles as shown in Table 3, 4 and 5 could also be attributed to difficulty in identifying hoax articles.

Model	Class	Precision	Recall	F1-score
LR	*	41.57	44.32	41.91
	S	51.06	39.41	44.48
	H	21.17	12.77	15.93
	P	38.79	47.06	42.53
	T	55.27	78.03	64.71
CNN	*	41.36	44.25	39.79
	S	53.67	30.49	38.89
	H	20.88	6.80	10.26
	P	38.74	64.40	48.37
	T	52.17	75.31	61.64
Bi-LSTM	*	43.36	46.18	43.76
	S	48.10	46.78	47.43
	H	25.65	12.09	16.43
	P	38.28	51.51	43.92
	T	61.41	74.34	67.26
DistilBERT	*	43.33	42.73	37.86
	S	46.62	44.18	45.37
	H	18.32	6.71	9.83
	P	68.01	28.54	40.21
	T	40.37	91.49	56.02

Table 4: Cross dataset 4-way classification scores (Synonym Augmentation)

For example, in Table 3, Logistic Regression performed the worst at classifying hoax articles with a 7.20 F1-score compared to 59.46, 48.96 and 63.68 F1-scores for satire, propaganda and trusted news respectively, even when the news articles were processed and cleaned in the same way.

Hoax news articles may be difficult to classify due to the nature of hoaxes being written similarly to legitimate news sources. In the first place, hoax news articles are written to deceive the public, and given that content in the news articles cannot be fact-checked beforehand by the model, hoax news articles with blatant misinformation are harder to identify.

5.3 Classification of Trusted news articles

In Table 3, we see that Logistic Regression (63.68), CNN (53.88), Bi-LSTM (67.93) and DistilBERT (48.99) performed better when classifying trusted news articles compared to non-trusted news articles. We interpret these results as the model learning the classification between trusted and non-trusted news articles better than learning boundaries between non-trusted news articles.

We posit that this is due to the appropriate feature extraction in the models. In Logistic Regression, features such as the Flesch Kincaid Reading Ease and *tf-idf* help classify between trusted and non-trusted news. Similarly, CNN and DistilBERT use word embeddings and sentence level representations that is different depending on the trustworthiness of news article.

Model	Class	Precision	Recall	F1-score
LR	*	41.97	48.34	42.83
	S	57.62	60.26	58.91
	H	16.67	1.97	3.50
	P	41.76	65.91	51.12
	T	51.85	65.20	57.76
CNN	*	37.66	41.14	38.29
	S	40.49	44.09	42.21
	H	20.11	7.46	10.88
	P	38.61	53.11	44.71
	T	51.47	59.91	55.36
Bi-LSTM	*	44.31	47.21	48.45
	S	50.00	46.98	48.45
	H	26.86	11.89	16.48
	P	38.72	53.09	44.77
	T	61.67	76.89	68.45
DistilBERT	*	39.55	44.20	39.19
	S	45.31	45.10	45.20
	H	22.66	3.94	6.72
	P	45.03	51.74	48.15
	T	45.22	76.03	56.71

Table 5: Cross dataset 4-way classification scores (NER Masking)

5.4 Comparison to Human Ceiling

According to Icard et al. (2024), humans generally show a moderate inter-annotator agreement of 62.0 for fake news labelling. This value is used as the human ceiling for evaluation. We generalise F1-score and compare it against the inter-annotator agreement value as a benchmark and gauge model performance. In Table 2, the Logistic Regression (73.93) and Bi-LSTM (73.77) performed better than the human ceiling, whereas CNN (59.55) and DistilBERT (53.17) performed close to it. We thus make the claim that our models are relevant in classifying fake news.

5.5 Synonym Augmentation

Synonym Augmentation improved the overall F1-scores for CNN and DistilBERT, but reduced the F1-scores for Logistic Regression and Bi-LSTM. This section discusses potential reasons for this outcome.

The overall F1-scores for CNN and DistilBERT improved by 3.94 and 4.03 respectively from the base results on the external dataset in Table 3 to the results with synonym augmentation in Table 4. This could be due to the feature extraction of these models, where CNN was able to capture spatial information about the word embeddings in its embedding layer, and DistilBERT capturing sentence representations. Synonyms tend to have a similar embedding vector. Thus, given that synonym augmentation slightly changes the representations for the augmented example in these 2 models, these models might better learn the boundaries to which a certain classification is made, thereby improving generalizability.

However, the F1-scores of Logistic Regression dropped by 2.91. We associate this drop with certain features such as *tf-idf* in the feature set, which causes 2 different words, despite their similar semantic meaning, to be classified in a manner mutually exclusive to each other as *tf-idf* does not capture similarity between words. Synonym augmentation also plays a minimal role in verb counting, since synonyms to verbs and adverbs are also likely to be verbs and adverbs themselves. Despite this, given that the number of features present in *tf-idf*, even after feature selection, is much larger than the other features, it would likely contribute to a lower F1-score.

5.6 NER masking

NER masking was intended to abstract named entities from news articles so that models are able to better capture the structural features and emotional wording of news articles rather than named entities when classifying text.

When comparing results in Table 3 and Table 5, we see that CNN, DistilBERT and Bi-LSTM have improved overall F1-scores to differing extents. It is interesting to note that the F1-scores for hoax news articles have gone down as a result of NER masking. From this, we can infer that hoax news articles rely on named entities and their context to a greater extent to classify correctly, whereas satire, propaganda and trusted news sources do not do so to the same extent.

However, Logistic Regression decreased in F1-scores when trained and evaluated on the NER masked train corpus and test corpus 2 respectively. This is likely due to named entities being associated with the classification of a sentence as satire, hoax, propaganda or trusted news in the selected features. For Logistic Regression, named entities that appear in the example are likely to be flagged as important words or phrases (unigrams or bigrams) by *tf-idf*. Removing named entities in such a case would result in a lower F1-score.

An additional example taken from test corpus 2 further shows the effects of NER masking. In this example, we replace a satire news article about Donald Trump and Harland Dorrinson with <PERSON>. We used the DistilBERT in this example to predict the label of the masked news article. Ini-

tially, the model predicted that this example is a hoax, which is false. However, after named entities has been removed, the model predicted that the example is satire, which is correct. This shows how named entities might skew the classification of a test example to a certain label even when content itself may not be that of a hoax article.

6 Conclusion

In this paper, we explore the generalization of fine-grained fake news detection and classification with the SHPT taxonomy, using the proposed synonym augmentation and NER masking methods. We have explored to what extent these methods help our proposed LR, CNN, Bi-LSTM and DistilBERT models generalize on unseen news articles, and discussed why these methods may have differing effects on generalization performance based on the design of each model.

Of our proposed approaches, synonym augmentation, which expands the sample space with slightly different but semantically similar representations of existing samples, seems to interact well with our CNN and DistilBert models with an overall increase in F1-scores, while our LR and Bi-LSTM models experienced an overall decrease in F1-score, though of a smaller magnitude. We claim that these differences arise from how synonym augmentation affects the different features captured by each model extracted to varying extents.

NER masking had a slightly different impact on some models, with CNN, Bi-LSTM and DistilBERT models seeing an increase in generalization performance, with a 2.92, 5.36 and 2.44 increase in F1-score respectively, and LR seeing a 1.99 drop in F1-score. DistilBERT seems to have had a change in performance of a greater magnitude than other models, suggesting that the emphasis on structural features with the abstraction of named entities benefits from its learned representation for classification tasks to a greater extent, relative to the features captured by other models.

6.1 Future Direction

We propose possible directions future exploration into the generalization of fake news classification can take.

6.1.1 NER Masking on Unseen Data with Different Contexts

We have shown in our results that named entities influence fake news classification, evidenced by the differing F1-scores when NER masking is applied. This could be attributed to most of the test examples in test corpus 2 and the train corpus using news articles from the United States, allowing named entities to play a larger role in classifying fake news. However, NER masking has yet to be tested on a test dataset where named entities between the training and testing datasets are mutually exclusive. This might serve as a potential research topic in the future.

6.1.2 Confidence scoring

Another possible direction is confidence scoring. Confidence scoring is a metric that ensures models are sufficiently confident about their classification. Calculation of the confidence score is based on maximum probability estimates of labels for all test examples. Given that X is the set of testing examples:

$$\forall x \in X, Confidence_x = max_{l \in labels} P(l)$$

Each model is used to predict the labels in the test set. A label is only accepted if the testing prediction falls into the top 80% of the prediction probabilities for the set of test examples in a single model. For training predictions that consistently fall into the bottom 20%, predict the test examples using an ensemble model detailed in the following section. Tied predictions in the top 80% also tie-break using the ensemble model. These thresholds are motivated by the testing of probability predictions using different feature sets in LR, where a sharp drop in prediction probabilities was observed in the lowest 20% of prediction probabilities in the testing set regardless of features used.[9]

6.1.3 Ensemble bagging

Research has shown that ensemble methods improve the accuracy of prediction compared to standalone models (Zhou, 2012). Expanding on this idea, a bagging of 4 models - LR, CNN, Bi-LSTM, DistilBERT to form an ensemble for classification of news could be explored. The ensemble model could use different techniques like ensemble bagging and stacking in order to achieve a better generalization performance compared to standalone models.

[9] A.2, Figure 1

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1:e9.

Carlos Carrasco-Farré. 2022. The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, 9:162.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763, Online. Association for Computational Linguistics.

Benjamin Icard, François Maine, Morgane Casanova, Géraud Faye, Julien Chanson, Guillaume Gadek, Ghislain Atemezing, François Bancilhon, and Paul Égré. 2024. A multi-label dataset of french fake news: Human and machine insights.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882.

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.

Arjun Roy, Kingshuk Basak, Asif Ekbal, and Pushpak Bhattacharyya. 2018. A deep ensemble framework for fake news detection and classification.

Roney Santos, Gabriela Pedro, Sidney Leal, Oto Vale, Thiago Pardo, Kalina Bontcheva, and Carolina Scarton. 2020. Measuring the impact of readability features in fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1404–1413, Marseille, France. European Language Resources Association.

Vaibhav Vaibhav, Raghuram Mandyam Annasamy, and Eduard Hovy. 2019. Do sentence interactions matter? leveraging sentence level representations for fake news classification.

Hongren Wang. 2023. Multi-label text classification using glove and neural network models.

Liqiang Wang, Yafang Wang, Gerard de Melo, and Gerhard Weikum. 2018. Five shades of untruth: Finer-grained classification of fake news. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 593–594.

Zhi-Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms*, volume 14. CRC press.

Acknowledgements

We would like to thank the CS4248 teaching team, and our mentor, Xian He, for their guidance and assistance in this project.

Statement of Independent Work

1A. Declaration of Original Work. By entering our Student IDs below, we certify that we completed our assignment independently of all others (except where sanctioned during in-class sessions), obeying the class policy outlined in the introductory lecture. In particular, we are allowed to discuss the problems and solutions in this assignment, but have waited at least 30 minutes by doing other activities unrelated to class before attempting to complete or modify our answers as per the class policy.

We have documented our use of AI tools (if applicable) in a following table, as suggested in the NUS AI Tools policy[10]. This particular document did not use any AI Tools to proofcheck and was constructed and edited purely by manual work.

Signed, A0216121Y, A0217701N, A0218441L, A0230337N, A0235341N, A0235393B

A Appendix

A.1 Individual Contributions

A.1.1 A0216121Y

- -

A.1.2 A0217701N

- Preliminary LR modelling
- Preliminary Naive Bayes modelling
- Preliminary formatting of fulltrain.csv from LUN

A.1.3 A0218441L

- External dataset sourcing - Hoax, Propaganda
- CNN modeling
- Report — **4.2.2**

[10] https://libguides.nus.edu.sg/new2nus/acadintegrity, tab "AI Tools: Guidelines on Use in Academic Work"

A.1.4 A0230337N

- External dataset sourcing - Satire, Trusted
- Preprocessing
- Feature extraction
- Synonym augmentation implementation
- NER masking implementation
- LR modelling
- Report — **1**, **3**, **4.2.1**, **5**, **6.1**

A.1.5 A0235341N

- Preprocessing
- DistilBERT modelling
- Report — **Abstract**, **2**, **3**, **4**, **6**

A.1.6 A0235393B

- BiLSTM modelling
- Report — **4.2.3, 4.3, 4.4**

A.2 Plots

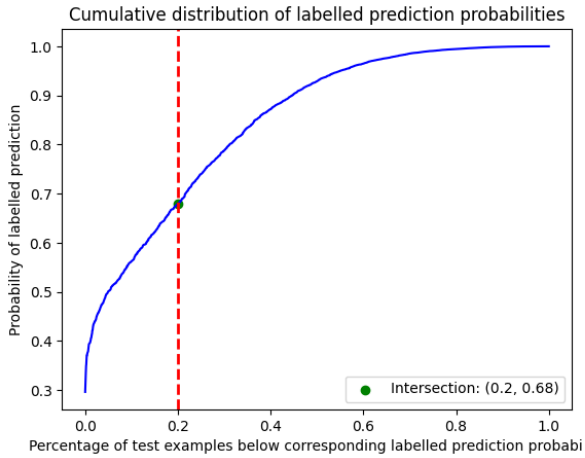

Figure 1: Cumulative distribution of labelled prediction probabilities