Transforming Fake News: Robust Generalisable News Classification Using Transformers

Ciara Blackledge
School of Computing, Newcastle University
Newcastle Upon Tyne, United Kingdom
c.blackledge1@newcastle.ac.uk

Amir Atapour-Abarghouei
Department of Computer Science, Durham University
Durham, United Kingdom
amir.atapour-abarghouei@durham.ac.uk

***Abstract*—As online news has become increasingly popular and fake news increasingly prevalent, the ability to audit the veracity of online news content has become more important than ever. Such a task represents a binary classification challenge, for which transformers have achieved state-of-the-art results. Using the publicly available ISOT and Combined Corpus datasets, this study explores transformers' abilities to identify fake news, with particular attention given to investigating generalisation to unseen datasets with varying styles, topics and class distributions. Moreover, we explore the idea that opinion-based news articles cannot be classified as real or fake due to their subjective nature and often sensationalised language, and propose a novel two-step classification pipeline to remove such articles from both model training and the final deployed inference system. Experiments over the ISOT and Combined Corpus datasets show that transformers achieve an increase in F_1 scores of up to 4.9% for out of distribution generalisation compared to baseline approaches, with a further increase of 10.1% following the implementation of our two-step classification pipeline. To the best of our knowledge, this study is the first to investigate generalisation of transformers in this context.**

***Index Terms*—Fake News Detection, Transformers, Natural Language Processing, Deep Learning**

I. Introduction

It has been argued that one of the great benefits of internet technology "is that it places a powerful tool of communication in the hands of the people" [12]. This democratisation of communication has allowed a shift in the way that news is written, shared and read, with the classical "top-down communication between elites and the general public" now being "subverted by horizontal communication between citizens through the internet" [12].

According to Ofcom's 2020 news consumption report [42], 65% of adults use the internet as a news platform, while in 2016 this proportion was 41% [41], showing a steep increase in the availability and popularity of news online. Fletcher and Parker propose that this widespread exposure to a broad spectrum of news sources has created "a more pressing need to filter credible information" [18].The issue of mistrust in online news has become prevalent in recent years, with the term "fake news" being coined in 2016 as a response to the flurry of misinformation spread surrounding the 2016 presidential election [17]. Consequently, with so many now accessing news online, verifying and auditing its veracity has become increasingly important.

Much of the research into fake news to date falls into two categories: examining solely textual content [59] and investigating mainly social context [5], [51]. Zhou et al. take a linguistic approach to the problem, proposing a theory driven model that explores content at "lexicon-level, syntax-level, semantic-level, and discourse-levels" [59]. Conversely, Shu et al. pay particular attention to social context and examine the "tri-relationship, the relationship among publishers, news pieces, and users" [51], determining that there are correlations between these factors and the likelihood of a news story being fake, while Albahar gives equal emphasis to news content and user comments [5]. Although this social context may improve the accuracy of news classification, such approaches limit detection to the point at which misinformation has already been posted, read and shared, hence only partially offering a solution to the fake news problem. This study instead focuses on identifying fake news solely from its textual content.

Transformers have reached state-of-the-art capabilities on a range of natural language processing tasks through "relying entirely on an attention mechanism to draw global dependencies between input and output" [55]. They have thus "rapidly become the dominant architecture for natural language processing" [57]. There is an increasing body of work exploring transformers for news classification, particularly in the context of politics (following the 2016 US presidential election) and public health (following the surge in misinformation relating to the COVID-19 pandemic) which have shown promising results when testing in distribution generalisation. However, little research has been carried out investigating the out of distribution generalisation abilities of transformers in this context.

Moreover, the term "fake news" and what it constitutes is widely contested, particularly in political contexts [22], [36]. As a result, it is likely that dataset labelling varies depending on the interpretation of fact-checkers, making these labels subjective and inconsistent. Deep learning models can be "prone to learning spurious correlations and memorizing high-frequency patterns" within data that do not generalise [50] and so these unreliable labels may harm generalisation performance. News articles that contain a large proportion of subjective information, and are therefore opinion-based, show a low consensus among fact-checkers [37] and so may make up a significant portion of potentially mislabelled samples.

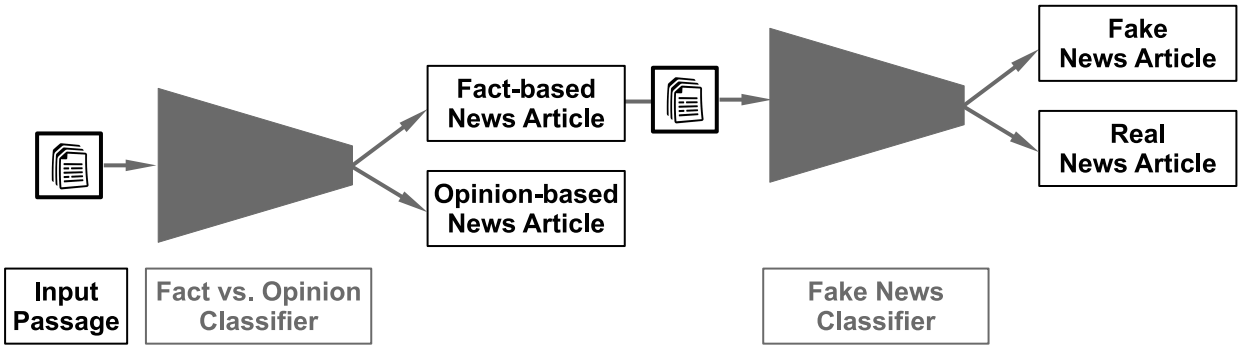

Fig. 1: Two-step opinion filtering pipeline proposed for news classification.

In order to mitigate this unreliable labelling, we propose a novel two-step classification pipeline that identifies and removes opinion-based news articles from the data used to train the final news classification model. In the proposed two-step system, the input new article is first classified as either a fact-based or opinion-based news article. Fact-based news articles are subsequently passed into a downstream classifier, which identifies whether the news article is real or fake. Fig. 1 shows the steps involved in this classification pipeline. Further details of the proposed approach is provided in Section IV. To enable reproducibility, an implementation of the proposed approach is publicly available[1].

In order to assess the contributions of this work, we provide an overview of previous related work in both the fields of text classification and fake news detection in Section II.

II. Related Work

We consider prior work in the context of text classification (Section II-A) and fake news detection (Section II-B).

A. Text Classification

Text classification is one of the core tasks encompassed by natural language processing, aiming to assign a document to predefined classes [13]. While studies have shown machine learning techniques to achieve high performances in this field [3], [7], [24], [43], the performance of such methods depends heavily on the data representation they are given.

However, deep learning models are able to represent the world as a "nested hierarchy of concepts" through hidden layers of learning [21]. Neural networks can therefore "transform low-level features of the data into high-level abstract features" and so are generally "stronger than shallow machine learning models in feature representation" [16]. Deng and Liu suggest [14] that while machine learning approaches to natural language processing tasks have reached relatively high levels of performance, they still fall short of human abilities, largely due to this "bottleneck" of feature engineering [14].

Minaee et al. provide a comprehensive review of transformers, machine learning and deep learning methods for text classification and find that deep learning leads to "significant improvements" across all tasks carried out [39]. Gonzalez-Carvajal and Garrido-Merchan also show BERT based transformers to outperform traditional machine learning techniques on a range of text classification tasks, noting in particular the importance of transfer learning, attributing their strong results to pre-training [20].

In the context of fake news classification, which is a complex challenge even for experienced fact-checkers, models capable of understanding the "subtleties involved in conveying messages through text" [53] are necessary to tackle this nuanced problem, with transformers achieving the best results to date on such tasks.

B. Fake News Detection

The consensus among researchers in the field is that there is little consensus regarding the exact definition of fake news, with Lilleker stating that this phrase has become "a catch-all term with multiple definitions" [36]. As a result of this ambiguously defined problem, there has been a historic lack of labelled benchmark datasets which has "dramatically limited" statistical approaches to the fake news problem [56]. Consequently, Wang created the popular LIAR benchmark dataset, containing 12k rows of short texts labelled as one of five classes ranging from "True" to "Pants on fire". Following this, a selection of datasets large enough for training deep learning models have been created in recent years, with Ahmed et al. releasing the ISOT fake news dataset in 2018 [3], and most recently, Khan et al. releasing the Combined Corpus dataset [32] in 2021.

Ahmed et al. provide initial benchmark results on the ISOT dataset showing an accuracy score of 92% using unigram features and a Linear SVM classifier [3]. In another work, Ahmad et al. propose an ensemble machine learning approach of combining Decision Tree, Random Forest and Extra Tree Classifiers using bagging to aggregate their outputs and increase this accuracy to 99.8% [2].

Alameri directly compares the performance of machine learning and deep learning text classifiers on this ISOT dataset, with the Long Short Term Memory (LSTM) model outperforming others, achieving the highest performance across accuracy, precision, recall and F_1 metrics [4]. Nasir et al. propose a hybrid Convolutional Neural Network (CNN) and Recurrent

[1] https://github.com/CiaraBee/fake_news_classification

Dataset	Total Rows of Data	Fake News	Real News
ISOT Fake News Dataset	44,898	21,417	23,481
Combined Corpus Dataset	79,548	40,689	38,859

TABLE I: Details of the ISOT and Combined Corpus datasets.

Neural Network model (RNN) that "makes use of the ability of the convolutional neural networks to extract local features and of the LSTM to learn long-term dependencies", and achieve 99% accuracy on the ISOT dataset [40], reinforcing the LSTM as a strong candidate for this task.

Aggarwal et al. investigate the performance of pre-trained BERT models on fake news detection and find that the "BERT model considerably outperforms other approaches even with minimal to no engineering of features", concluding that transfer learning "can yield good results in the case of detection of fake news" [1]. Radford et al. demonstrate that pre-training contributes to the strong performance of transformers on a wide variety of natural language processing fields, showing in ablation studies that its removal "hurts performance across all the tasks" [46], and reinforcing BERT as the state-of-the-art in this context due to its "deep understanding of the language", which is considered "necessary to detect the subtle stylistics differences in the writing of the fake articles" [6].

Khan et al. [32] offer a benchmark study of fake news detection and find that "pre-trained BERT based models outperform the other models not only on the overall datasets but also on smaller samples". This capacity for strong performance even on small datasets provides evidence that transformers are less prone to overfitting in this context, indicating that they may generalise well.

With vast amounts of news content constantly being published online, the ability of any fake news detection model to adapt to unseen data is critical in tackling this problem. Nasir et al. investigate generalisation to unseen news articles from a deep learning approach using their hybrid CNN-RNN model, training on the ISOT fake news dataset and testing on another, which results in "poor generalisation", achieving results that "indicate overfitting" [40]. However, to the best of our knowledge there has been no research on the out of distribution generalisation abilities of transformers in this context. This study, therefore, aims to fill this gap by investigating the generalisation capabilities of transformers and proposing a two-step classification pipeline to enhance their overall performance.

III. Data

For the purposes of this study, which aims to investigate the applications of deep learning on the fake news problem using solely linguistic features, a large text focused dataset is needed. Marcus notes that deep learning is "data hungry" and "works best when there are thousands, millions or even billions of training examples" [38]. The following two datasets have therefore been identified as appropriate for this project due to their size and content. Table I details the contents of these datasets, showing the total number of articles and proportions of real and fake news present. This table shows that the data is largely balanced, with the ISOT dataset (ISOT) containing slightly more fake news, and the Combined Corpus dataset (CC) containing slightly more real news. Further details of these two datasets are outlined in the following.

A. ISOT Fake News Dataset

In 2017, Ahmed et al. introduced the ISOT fake news dataset [3] containing 45k full length articles from real world sources, with real articles collected from Reuters.com and fake articles collected from various unreliable sources. Fake news articles in this dataset were identified and sourced from Politifact.com, a not-for-profit national news organization that uses human fact-checkers to identify fake news, largely focusing on political news. This dataset therefore mainly contains news articles relating to the 2016 US presidential election and has been widely used in fake news detection studies in this field [3], [4], [24], [40].

B. Combined Corpus Dataset

In 2021, Khan et al. introduced the Combined Corpus dataset (CC) [32] containing nearly 80k rows of data, with 51% being real news and 49% being fake news. In contrast to ISOT, the creators of this dataset actively sought out news from a wide variety of sources covering a range of topics including "national and international politics, economy, investigation, health-care, sports, entertainment, and others" [32], making this dataset larger in both size and scope. This dataset spans articles from 2015 to 2017 and covers a wide range of fake news types such as "hoax, satire, and propaganda" [32]. To the best of our knowledge, this is at present the largest publicly available fake news dataset.

C. Data Preprocessing

While this work primarily focuses on utilising solely textual features for news classification and so no feature engineering is carried out, a number of preprocessing steps have been taken to clean and prepare the ISOT and CC datasets for use. In order to preserve as many characteristic textual features as possible, all spelling and grammatical errors present within each dataset have been maintained. The focus of data preprocessing is therefore that of removing extraneous elements that may detract from these textual features.

In this vein, all URLs, punctuation, IP addresses and links within the body of the text have been removed from the

datasets via regular expressions [54]. Stopwords are subsequently removed from the cleaned text [8], which is then vectorised for the LSTM model [10] and the conventional ML models [45] used as baselines in this work. The HuggingFace library [57] provides model-specific tokenizers for each transformer implementation. After completing these preprocessing steps, the data is ready for input into each model considered.

IV. Proposed Approach

In order to assess the efficacy of transformers on news classification, we propose an approach comparing their performance to baseline machine and deep learning approaches. To further explain this proposed approach, details of all models implemented are outlined below, with particular attention paid to the BERT based transformers and the differences between the three architectures considered. Moreover, further detail is given regarding the two-step classification pipeline introduced in this study to further expand on the rationale and implementation for doing so.

A. Baseline Models

Both machine and deep learning approaches are considered for baseline models with Logistic regression [11], Naïve Bayes [25] and Random Forest classifiers [29] forming the machine learning baseline comparison.

Alameri identifies LSTMs [30] to be the best performing deep learning model for the task of news classification [4] and so an LSTM will therefore be used as the deep learning benchmark for this work.

BERT (Bidirectional encoder representation from transformers) [15] builds upon the original transformer architecture [55] and has reached new state-of-the-art capabilities on a variety of text classification tasks [15], [35], [52] by learning from a combination of pre-training and fine tuning. BERT is therefore considered "a must-have baseline" for natural language processing tasks [48] and so is examined for its use on fake news detection in this study along with two other BERT based model types: DistilBERT [49] and deBERTa [26].

B. BERT

BERT models [15] differ from the original transformer [55] by allowing bidirectional language understanding. BERT models undergo unsupervised pretraining in a two-step process made up of masked language modelling and next sentence prediction. During masked language modelling, input tokens are randomly masked and subsequently predicted in order to obtain a "deep bidirectional representation" [15]. This allows BERT to counter the "unidirectional constraint" [19] of other language models such as GPT [46] by not allowing the model to "see itself" and thus "trivially predict the next token" when learning both right to left and left to right [19]. The next stage of pretraining takes the form of binarised next sentence prediction where sentence A precedes sentence B 50% of the time, allowing the model to learn the "relationship between two sentences" [19]. BERT models are then fine tuned by adding a classification layer and updating all parameters based on a downstream task, in this case, fake news classification.

C. DistilBERT

DistilBERT is a distilled BERT architecture 40% smaller than its predecessor and capable of achieving similar results while being 60% faster during inference [49]. This lightweight model is obtained through distillation, in which knowledge is transferred from a large model to a smaller, more compact counterpart [28]. DistilBERT has been shown to retain 97% of the natural language understanding capabilities of its larger equivalent [49].

D. DeBERTa

Building upon BERT, DeBERTa (Decoding-enhanced BERT with disentangled attention) proposes the addition of "a disentangled attention mechanism and an enhanced mask decoder" [26]. Dissimilarly to BERT, DeBERTa word encodings are made up of two vectors that encode both content and relative position using disentangled matrices. To enhance the masked language modelling phase of pre-training, DeBERTa "incorporates absolute word position embeddings" right before the model decodes the masked words [26].

E. Fact vs. Opinion Classification

As discussed in Section I, an important contribution of this work is the introduction of an additional step in the overall fake news classification pipeline to identify and remove opinion-based passages, providing a cleaner and more accurate process for detecting fake news articles.

Lim investigates fact-checking for political news and finds that the "the rate of agreement on its factual accuracy is quite low for statements in the relatively ambiguous scoring range" [37], indicating that there is a large amount of uncertainty in classifying claims that do not directly confirm or contradict a fact, i.e., subjective or unclear claims. Furthermore, Graves identifies that many fact-checking organisations adopt the rule that there is "no way to check a statement of opinion" [22], confirming that opinions cannot be fact-checked and thus exist outside the scope of news classification.

In accordance with this hypothesis that opinions therefore cannot be labelled as real or fake, opinion-based or highly subjective news articles present in the two datasets used in this work could therefore be considered mislabelled samples. These mislabelled samples pose the risk that models may learn incorrect patterns in the data, which would impair, in particular, their ability to generalise [31]. Lallich et al. find that removing mislabeled classes from training data improved classification on a range of datasets and so this study therefore explores the effect of implementing an additional classification step in which opinion-based articles are identified and removed from the data, creating a filtered training dataset [34].

Moreover, similarly to opinion-based news stories, fake news articles often use emotive and sensationalised language [51]. While this language is more commonly found in fake news, it is not exclusive to this class and so by removing opinion-based articles we aim to prevent models unintentionally learning to classify articles based on their degree of sensationalism.

While a number of studies have been carried out attempting to score sentence subjectivity using datasets of labelled short sentences [47], [58], to date no datasets have been found collating fact and opinion-based news stories, which tend to be significantly longer. A new dataset has therefore been created as part of this work, containing 50 rows of data taken from online news sources, of which 25 are factual and 25 are opinion-based. This dataset is publicly available[2].

Out of the three transformer architectures implemented in this work, DistilBERT achieved the best results on the task of fact vs opinion classification with an accuracy of 78% on this small dataset. This model is implemented prior to fake news classification, with articles identified as opinion-based then removed from the training dataset.

F. Methods

To assess the performance of the models outlined above on fake news classification, four key metrics have been recorded: accuracy, precision, recall and the F_1 score. Each baseline and transformer model has been trained on 80% of each dataset, with 10% used for validation and 10% reserved for testing. All transformer implementation is done in *PyTorch* [44], with AdamW [33] providing the best optimization ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$). Experimentation has been carried out using a Tesla P100 GPU with an Intel(R) Xeon(R) CPU @ 2.00GHz processor [9].

V. Experimental Results

Experiments have been run using the models outlined in section IV, testing both in distribution and out of distribution generalisation by evaluating each model's performance on both holdout test sets with similar distributions to training data, and completely unseen datasets with varying topics, styles and class distributions. Following this, the two-step classification pipeline introduced in this study is implemented to compare generalisation performance with and without this additional step and assess its suitability and efficacy to this use case.

A. In Distribution Generalisation

The baseline and BERT based models have each been trained on the ISOT and CC datasets and evaluated on their holdout test sets to allow comparison between the performance of machine learning, deep learning and transformer based approaches. Table II shows the results of training and testing on the ISOT dataset, reporting accuracy, precision, recall and F_1 scores and Table III shows these evaluation metrics when trained and tested on the CC dataset.

Table II and III show that the three BERT based models outperform the baseline models across all evaluation metrics, with deBERTa scoring the highest across both datasets. The results for BERT and DistilBERT are very similar, with BERT marginally outperforming its more compact counterpart on the ISOT dataset, and DistilBERT performing best on the CC dataset. This indicates that, while smaller, DistilBERT retains much of the natural language understanding of its larger equivalent, performing equally as well in this context but at a much faster speed (60% faster at inference [49]), making it a strong candidate for the task of news classification. These results clearly indicate that transformers are able to learn meaningful patterns and generalise well to datasets with similar distributions. However, with large amounts of online news, including fake news, being published every day by a wide range of sources, it is important to test the robustness of these models when faced with data of varying content, length and writing style.

[2]https://github.com/CiaraBee/fake_news_classification

Model	ISOT Dataset			
	Accuracy	Precision	Recall	F_1 Score
Logistic Regression	0.977	0.977	0.976	0.977
Naïve Bayes	0.939	0.938	0.940	0.939
Random Forest	0.957	0.959	0.954	0.956
LSTM	0.969	0.969	0.969	0.969
DistilBERT	0.989	0.989	0.989	0.989
BERT	0.990	0.990	0.989	0.989
deBERTa	**0.997**	**0.997**	**0.997**	**0.989**

TABLE II: Comparison results of baseline and BERT based transformer models when trained on 80% of the ISOT dataset and tested on 10%.

Model	Combined Corpus Dataset			
	Accuracy	Precision	Recall	F_1 Score
Logistic Regression	0.962	0.961	0.962	0.962
Naïve Bayes	0.866	0.871	0.864	0.865
Random Forest	0.927	0.927	0.926	0.926
LSTM	0.942	0.944	0.941	0.942
DistilBERT	0.983	0.983	0.983	0.983
BERT	0.980	0.980	0.981	0.981
deBERTa	**0.988**	**0.988**	**0.988**	**0.988**

TABLE III: Comparison results of baseline and BERT based transformer models when trained on 80% of the CC dataset and tested on 10%.

B. Out of Distribution Generalisation

In order to assess whether transformers are able to generalise to unseen data with different distributions, experiments have been run in which each model has been trained on either the ISOT or CC dataset and then tested on the other. The results of these experiments are reported in Tables IV and V.

Tables IV and V show a clear drop in performance when compared with the results of tables II and III. However, as these datasets are from different sources, this is to be expected since a difference in the data distribution between the training and test sets always leads to a drop in performance. Nonetheless, these results show that BERT based models consistently achieve better performance than the baseline comparisons across all evaluation metrics. A major advantage of transformers for this task is their pre-training on large and varied datasets which Hendrycks et al. suggest improves robustness and generalisation [27].

Model	Trained on ISOT Dataset			
	Accuracy	Precision	Recall	F_1 Score
Logistic Regression	0.641	0.678	0.649	0.612
Naïve Bayes	0.642	0.689	0.647	0.623
Random Forest	0.645	0.679	0.649	0.631
LSTM	0.632	0.746	0.639	0.691
DistilBERT	0.670	0.716	0.674	0.655
BERT	0.670	**0.748**	**0.715**	**0.702**
deBERTa	**0.697**	0.702	0.699	0.695

TABLE IV: Comparison results of baseline and BERT based transformer models when trained on the ISOT dataset and tested on the Combined Corpus dataset.

Model	Trained on CC Dataset			
	Accuracy	Precision	Recall	F_1 Score
Logistic Regression	0.726	0.789	0.736	0.737
Naïve Bayes	0.665	0.680	0.642	0.635
Random Forest	0.707	0.721	0.689	0.688
LSTM	0.681	0.754	0.649	0.629
DistilBERT	**0.775**	**0.846**	**0.751**	**0.750**
BERT	0.670	0.755	0.637	0.610
deBERTa	0.730	0.800	0.703	0.695

TABLE V: Comparison results of baseline and BERT based transformer models when trained on the Combined Corpus dataset and tested on the ISOT dataset.

Table V shows a generally stronger generalisation performance across models, potentially due to the larger size and broader topic scope of the CC dataset. Moreover, these results show slightly more variation across datasets and models, with the logistic regression classifier achieving higher recall and F_1 scores than BERT and deBERTa.

While the transformers perform well overall on this task, it is clear that there are more incorrect predictions being made compared to in distribution generalisation. As the values reported in Tables VI and VII are macro averages, we can better understand these results by investigating the evaluation metrics for each class.

These results show that while the BERT based transformers generalise well overall, there are large differences in the precision and recall values for each class, indicating skews towards predicting a certain class. Of particular note are the evaluation metrics below 0.5, which indicate performance lower than random chance and suggest that these models may be learning spurious connections from the data that significantly hinder generalisation to unseen data. In order to investigate whether mislabelled subjective samples may be the cause of such incorrect patterns, we go on to assess the effect of implementing the two-stage classification pipeline demonstrated in Figure 1.

1) Filtering Opinion-Based Articles: As outlined in Figure 1, a filtered training dataset has been created. All baseline and BERT based transformers have been trained on this filtered training data and evaluated against the alternative dataset to investigate how this additional step affects out of distribution generalisation. The results from the two-step classification pipeline are compared with the one-step classification in Tables VIII and IX below.

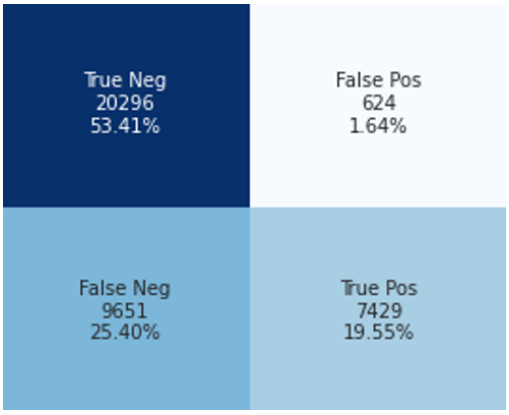

Fig. 2: Confusion matrix for deBERTa model trained on CC dataset and tested on ISOT.

Tables VIII and IX clearly show a difference in the effectiveness of removing opinion-based articles between datasets, with models trained on the CC dataset seeing improvements from their removal, while those trained on the ISOT dataset overall do not. It is important to note that in removing opinion-based articles from the data, the size of the resulting dataset is reduced. The LSTMs and transformers, being neural networks, have a large number of parameters and so "require large amounts of data for training in order for over-fit avoidance and better model generalisation" [23]. The results shown in Table VIII suggest that the resulting dataset after removing opinion-based articles is too small for these models to generalise well, resulting in generally reduced performances.

However, when training on the larger more diverse CC dataset, Table IX shows that the removal of opinion-based articles allows deBERTa to achieve the highest performance across all metrics. DeBERTa achieves particularly large gains in generalisation with this filtering step applied, increasing its F_1 score on this task by 10.1%. With the exception of BERT, this opinion filtering step notably increases the generalisation performance of both the transformers and the LSTM.

Figure 2 shows the confusion matrix for the deBERTa model trained on the CC dataset and evaluated on the ISOT dataset. These results show that this model is largely skewed towards predicting negative values, with a high proportion of false negative and a low proportion of true positives. Figure 3 shows the confusion matrix for this same model after implementing the two-step classification pipeline. These results show a smaller proportion of false negative values along with an increased proportion of true positives, indicating that removing opinion-based articles has resulted in better classification and generalisation performance.

Model	Accuracy		Precision		Recall		F_1 Score	
	Fake	Real	Fake	Real	Fake	Real	Fake	Real
DistilBERT	0.670	0.670	0.611	0.821	0.896	**0.453**	0.726	0.583
BERT	0.711	0.711	0.646	0.849	0.903	0.527	0.753	0.651
deBERTa	0.697	0.697	0.662	0.742	0.774	0.622	0.714	0.677

TABLE VI: Results of generalisation experiments for transformers trained on the ISOT dataset and tested on the CC dataset for each class.

Model	Accuracy		Precision		Recall		F_1 Score	
	Fake	Real	Fake	Real	Fake	Real	Fake	Real
DistilBERT	0.775	0.775	0.980	0.712	0.510	0.991	0.671	0.829
BERT	0.670	0.670	0.879	0.631	**0.308**	0.965	**0.456**	0.763
deBERTa	0.730	0.730	0.923	0.678	**0.435**	0.970	0.591	0.798

TABLE VII: Results of generalisation experiments for transformers trained on the CC dataset and tested on the ISOT dataset for each class.

Model	Accuracy		Precision		Recall		F_1 Score	
	Two Step	One Step	Two Step	One Step	Two Step	One Step	Two Step	One Step
Logistic Regression	0.662	0.641	0.747	0.678	0.669	0.649	0.636	0.612
Naïve Bayes	0.630	0.642	0.634	0.689	0.632	0.647	0.629	0.623
Random Forest	0.639	0.645	0.653	0.679	0.642	0.649	0.634	0.631
LSTM	0.632	0.632	0.670	0.746	0.637	0.639	0.616	0.691
DistilBERT	0.678	0.670	0.722	0.716	0.682	0.674	0.713	0.655
BERT	**0.710**	0.670	0.740	**0.748**	0.715	**0.715**	0.733	**0.702**
deBERTa	0.674	0.697	0.675	0.702	0.673	0.699	0.675	0.695

TABLE VIII: Comparison results showing generalisation performance of all models with the two step classification pipeline applied and without when trained on the ISOT dataset and tested on the CC dataset.

Model	Accuracy		Precision		Recall		F_1 Score	
	Two Step	One Step	Two Step	One Step	Two Step	One Step	Two Step	One Step
Logistic Regression	0.769	0.726	0.808	0.789	0.749	0.736	0.750	0.737
Naïve Bayes	0.631	0.665	0.757	0.680	0.591	0.642	0.535	0.635
Random Forest	0.694	0.707	0.721	0.721	0.672	0.689	0.667	0.688
LSTM	0.736	0.681	0.760	0.745	0.718	0.649	0.718	0.629
DistilBERT	0.714	0.681	0.735	0.739	0.729	0.652	0.713	0.693
BERT	0.667	0.670	0.629	0.755	0.699	0.638	0.638	0.672
deBERTa	**0.808**	0.730	**0.839**	0.800	**0.792**	0.703	**0.796**	0.695

TABLE IX: Comparison results showing generalisation performance of all models with the two step classification pipeline applied and without when trained on the CC dataset and tested on the ISOT dataset.

VI. Future Work

While the dataset used to train the fact/opinion classifier was very small (50 news articles), this additional step nonetheless resulted in notable increases in generalisation performance for models trained on the CC dataset, indicating that opinion filtering may aid the learning process by removing mislabelled samples. There is much scope for developing a larger dataset containing labelled fact and opinion-based news articles which would allow for better performance in identifying opinion-based articles, and thus likely improve generalisation further.

This study has shown that transformers may learn incorrect patterns from data that harm their out of distribution generalisation performance. Following this, there is further work to be done in quantifying, understanding and eventually preventing transformers learning spurious patterns in data. While this work explores removal of mislabelled data to tackle this problem, data augmentation has also been suggested as an approach to improve the robustness of models [50].

VII. Conclusion

The digitisation of media, and in particular social media, has allowed news, both real and fake, to propagate faster than ever before [17]. Auditing the veracity of news content posted online at the earliest point of detection possible is therefore crucial in tackling the fake news problem. Online

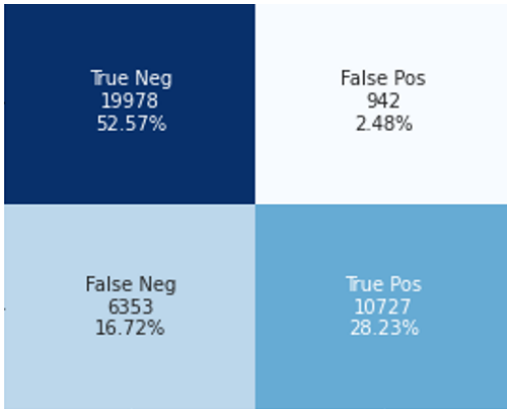

Fig. 3: Confusion matrix for deBERTa model trained on CC dataset and tested on ISOT with two-step pipeline implemented.

news as its core is simply text, and so this study assesses the effectiveness of using using transformers, the state-of-the-art in natural language processing, to classify news based solely on textual content, paying particular attention to out of distribution generalisation.

This study has shown that transformers such as BERT, DistilBERT and deBERTa outperform machine learning and deep learning baseline alternatives (logistic regression, naïve Bayes, random forest and LSTM classifiers) in news classification when testing in distribution generalisation and out of distribution generalisation, achieving a peak accuracy of 77.5% by the DistilBERT model on the latter. Additionally, we have addressed the subjective and inconsistent nature of fake news by proposing a two-step classification pipeline which identifies and removes opinion-based news articles from the training data used by the final news classifier. In doing so, the most subjective and therefore unpredictable samples are filtered out of the data to prevent models learning incorrect patterns that do not generalise. This two-step classification process improves the accuracy of deBERTa predictions by 7.8% to a peak of 80.8% and improves its F_1 score by 10.1%. However, the effectiveness of this method seems to be largely dependent on the dataset, with larger and more varied datasets producing superior results.

References

[1] Akshay Aggarwal, Aniruddha Chauhan, Deepika Kumar, Mamta Mittal, and Sharad Verma. Classification of fake news by fine-tuning deep bidirectional transformers based language model. page 163973, 04 2020.

[2] Iftikhar Ahmad, Muhammad Yousaf, Suhail Yousaf, and Muhammad Ovais Ahmad. Fake news detection using machine learning ensemble methods. *Complexity*, 2020, 2020.

[3] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer, 2017.

[4] Saeed Amer Alameri and Masnizah Mohd. Comparison of fake news detection using machine learning and deep learning techniques. In *2021 3rd International Cyber Resilience Conference (CRC)*, pages 1–6. IEEE, 2021.

[5] Marwan Albahar. A hybrid model for fake news detection: Leveraging news content and user comments in fake news. *IET Information Security*, 15(2):169–177, 2021.

[6] Wissam Antoun, Fady Baly, Rim Achour, Amir Hussein, and Hazem Hajj. State of the art models for fake news detection tasks. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 519–524, 2020.

[7] Amir Atapour-Abarghouei, Stephen Bonner, and Andrew Stephen McGough. Rank over class: The untapped potential of ranking in natural language processing. *arXiv preprint arXiv:2009.05160*, 2020.

[8] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[9] Ekaba Bisong. *Google Colaboratory*, pages 59–64. Apress, Berkeley, CA, 2019.

[10] François Chollet et al. Keras. https://keras.io, 2015.

[11] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[12] James Curran, Sharon Coen, Toril Aalberg, Kaori Hayashi, Paul K Jones, Sergio Splendore, Stylianos Papathanassopoulos, David Rowe, and Rod Tiffen. Internet revolution revisited: a comparative study of online news. *Media, Culture & Society*, 35(7):880–897, 2013.

[13] Mita K Dalal and Mukesh A Zaveri. Automatic text classification: a technical review. *International Journal of Computer Applications*, 28(2):37–40, 2011.

[14] Li Deng and Yang Liu. *Deep learning in natural language processing*. Springer, 2018.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Xuedan Du, Yinghao Cai, Shuo Wang, and Leijie Zhang. Overview of deep learning. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pages 159–164. IEEE, 2016.

[17] Álvaro Figueira and Luciana Oliveira. The current state of fake news: challenges and opportunities. *Procedia Computer Science*, 121:817–825, 2017.

[18] Richard Fletcher and Sora Park. The impact of trust in the news media on online news consumption and participation. *Digital journalism*, 5(10):1281–1299, 2017.

[19] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.

[20] Santiago González-Carvajal and Eduardo C Garrido-Merchán. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*, 2020.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[22] Lucas Graves. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique*, 10(3):518–537, 2017.

[23] Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*, 2019.

[24] Saqib Hakak, Mamoun Alazab, Suleman Khan, Thippa Reddy Gadekallu, Praveen Kumar Reddy Maddikunta, and Wazir Zada Khan. An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, 117:47–58, 2021.

[25] David J. Hand and Keming Yu. Idiot's bayes: Not so stupid after all? *International Statistical Review / Revue Internationale de Statistique*, 69(3):385–398, 2001.

[26] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[27] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[29] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[31] I Hsu, Ayush Jaiswal, Premkumar Natarajan, et al. Niesr: Nuisance invariant end-to-end speech recognition. *arXiv preprint arXiv:1907.03233*, 2019.

[32] Junaed Younus Khan, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. A benchmark study of machine learning models for online fake news detection. *Machine Learning with Applications*, 4:1–22, Apr 2021.

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learning Representations*, pages 1–15, 2014.

[34] Stéphane Lallich, Fabrice Muhlenbach, and Djamel A Zighed. Improving classification by removing or relabeling mislabeled instances. In *International Symposium on Methodologies for Intelligent Systems*, pages 5–15. Springer, 2002.

[35] Jieh-Sheng Lee and Jieh Hsiang. Patent classification by fine-tuning bert language model. *World Patent Information*, 61:101965, 2020.

[36] Darren Lilleker. Evidence to the culture, media and sport committee'fake news' inquiry presented by the faculty for media & communication, bournemouth university. 2017.

[37] Chloe Lim. Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848, 2018.

[38] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[39] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

[40] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. Fake news detection: A hybrid cnn-rnn based deep learning approach. *International Journal of Information Management Data Insights*, 1(1):100007, 2021.

[41] Ofcom. News consumption in the uk: 2016.

[42] Ofcom. News consumption in the uk: 2020.

[43] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

[44] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[46] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[47] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112, 2003.

[48] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[49] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[50] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34, 2021.

[51] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 312–320, 2019.

[52] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*, 2019.

[53] Aswini Thota, Priyanka Tilak, Simrat Ahluwalia, and Nibrat Lohia. Fake news detection: a deep learning approach. *SMU Data Science Review*, 1(3):10, 2018.

[54] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[56] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection, 2017.

[57] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[58] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136, 2003.

[59] Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.