Adapting Fake News Detection to the Era of Large Language Models

Anonymous ACL submission

Abstract

In the age of large language models (LLMs) and the widespread adoption of AI-driven content creation, the landscape of information dissemination has witnessed a paradigm shift. With the proliferation of both human-written and machine-generated real and fake news, robustly and effectively discerning the veracity of news articles has become an intricate challenge. While substantial research has been dedicated to fake news detection, this either assumes that all news articles are human-written or abruptly assumes that all machine-generated news is fake. Thus, a significant gap exists in understanding the interplay between machine-paraphrased real news, machine-generated fake news, human-written fake news, and human-written real news. In this paper, we study this gap by conducting a comprehensive evaluation of fake news detectors trained in various scenarios. Our primary objectives revolve around the following pivotal question: How can we adapt fake news detectors to the era of LLMs? Our experiments reveal an interesting pattern that detectors trained exclusively on human-written articles can indeed perform well at detecting machine-generated fake news, but not vice versa. Moreover, due to the bias of detectors against machine-generated texts (Su et al., 2023a), they should be trained on datasets with a lower machine-generated news ratio than the test set. Building on our findings, we provide a practical strategy for the development of robust fake news detectors. [1]

1 Introduction

Since Brexit and the 2016 US Presidential campaign, the proliferation of fake news has become a major societal concern (Martino et al., 2020). On the one hand, false information is easier to generate but harder to detect (Kumar and Shah, 2018). On the other hand, humans are often attracted to sensational information and spread it six times faster than truthful news (Vosoughi et al., 2018), which is a threat to both individuals and society as a whole.

Until recently, most online disinformation was human-written (Vargo et al., 2018), but recently a lot of it is AI-generated. With the continuing progress of natural language generation (Radford et al., 2019; Brown et al., 2020; Chowdhery et al., 2022), AI-generated content has become indistinguishable from human-written one, and it is also often perceived as more credible (Kreps et al., 2022) and trustworthy (Zellers et al., 2019; Spitale et al., 2023) than human-generated propaganda. This raises pressing concerns about the unprecedented scale of disinformation that AI models have enabled (Bommasani et al., 2021; Kreps et al., 2022; Buchanan et al., 2021; Goldstein et al., 2023).

While efforts to combat machine-generated fake news date back to as early as 2019 (Zellers et al., 2019), the majority of research in this field has primarily focused on detecting machine-generated text, rather than evaluating the factual accuracy of machine-generated news articles. In these studies, machine-generated text is considered to be always fake news, regardless of its content.

Previously, when generative AI was less prevalent, it was arguably reasonable to assume that most automatically generated news articles would be primarily used by malicious actors to craft fake news. However, with the remarkable advancement of generative AI in the last two years, and their integration in various aspects of our lives, these tools are now broadly adopted for legitimate purposes such as assisting journalists in content creation. Reputable news agencies, for instance, use AI to draft or to enhance their articles (Hanley and Durumeric, 2023). Nevertheless, the age-old problem of human-written fake news continues. This diverse blend of machine-generated genuine news, machine-generated fake articles, human-written fabrications, and human-written factual articles has shifted the way of news generation and the intricate

[1] Code and data would be released upon acceptance.

intermingling of content sources is likely to endure in the foreseeable future.

In order to adapt to the era of LLMs, the next generation of fake news detectors should be able to handle the mixed-content landscape of human/machine-generated real/fake news. While there exists a substantial body of research on fake news detection, it typically focuses exclusively on human-written fake news (Khattar et al., 2019; Kim et al., 2018; Paschalides et al., 2019; Horne and Adali, 2017; Pérez-Rosas et al., 2018) or on machine-generated fake news (Zellers et al., 2019; Goldstein et al., 2023; Zhou et al., 2023), essentially framing the problem as detection of machine-generated text. However, robust fake news detectors should primarily assess the authenticity of the news articles, rather than relying on other confounding factors, such as whether the article was machine-generated. Thus, there is a pressing need to understand fake news detectors on machine-paraphrased real news (**MR**), machine-generated fake news (**MF**), human-written fake news (**HF**), and human-written real news (**HR**).

Here, we bridge this gap by evaluating fake news detectors trained with varying proportions of machine-generated and human-written fake news. Our experiments yield the following key insights:

(1) Fake news detectors, when trained exclusively on human-written news articles (i.e., **HF**, **HR**), have the ability to detect machine-generated fake news. However, the reverse is not true. This observation suggests that, when the proportion of testing data is uncertain, it is advisable to train detectors solely on human-written real and fake news articles. Such detectors are still able to generalize effectively for detecting machine-generated news.

(2) Although the overall performance is mainly decided by the distribution of machine-generated and human-written fake news in the test dataset, the class-wise accuracy for our experiments suggests that, in order to achieve a balanced performance for all subclasses, we should train the detector on a dataset with a lower proportion of machine-generated news compared to the test set.

(3) Our experiments also reveal that fake news detectors are generally better at detecting machine-generated fake news (MF) than at identify human-written fake news (HF), even when exclusively trained on human-generated data (without seeing MF during the training). This underscores the inherent bias within fake news detectors (Su et al., 2023a). It is recommended to take these biases into consinderation when training fake news detectors.

Our contributions can be summarized as follows:

- We are the first to conduct comprehensive evaluation of fake news detectors across diverse scenarios where news articles exhibit a wide range of diversity, including both human-written and machine-generated real and fake content.
- Drawing from our experimental results, we offer valuable insights and practical guidelines for deploying fake news detectors in real-world contexts, ensuring that they remain effective amid the ever-evolving landscape of news generation.
- Our work lays the groundwork for understanding the data distribution shifts in fake news caused by LLMs, moving beyond simple fake news detection. We aim to heighten the research community's awareness of this evolving dynamic in human language and their larger impact.

2 Related Work

Fake news detection is the task of detecting potentially harmful news articles that make some false claims (Oshikawa et al., 2020). The conventional solution for detecting fake news is to ask professionals such as journalists to perform manual fact-checking (Shao et al., 2016), which is expensive and time-consuming. To reduce the time and the efforts for detecting fake news, researchers formulate this problem as a classification problem and seek solutions for automatic fake news detection from a machine learning perspective.

In general, there are two branches of the task formulation: one branch only consider human-written real vs. fake news, and the other one formulates this as detecting machine-generated text, thus automatically categorizing any machine-generated news as fake news.

2.1 Detecting Human-Written Real/Fake News

Before 2018, fake news were predominantly manually written (Vargo et al., 2018), which motivated early research on distinguishing human-written fake news from machine-generated ones. Various methods have been designed such as linguistic approaches (Chen et al., 2015; Rubin et al., 2016;

Pérez-Rosas et al., 2018), such as analysis of the writing style (Castelo et al., 2019) and of the content (Jin et al., 2016; VV and Zafarani, 2020); fact-checking approaches, which rely on the automatic verification of the claims made in news articles (Graves and Cherubini, 2016) or applying deep learning methods such as CNNs (Huang et al., 2017; He et al., 2016), LSTMs (Graves and Graves, 2012), or transformers (Devlin et al., 2019; Vaswani et al., 2017).

2.2 Distinguishing Machine-Generated from Human-Written News

With recent progress of natural language text generation (Radford et al., 2018, 2019; Zhao et al., 2023), there have also been rising concerns that malicious actors might generate fake news automatically using controlled generation (Mitchell et al., 2023; Zellers et al., 2019). To understand and to respond to neural fake news, (Zellers et al., 2019) studied the potential risk of neural disinformation and presented a model for neural fake news generation called GROVER, which allows for controlled generation of an entire news article. They generated fake news articles using GROVER, and experimented with distinguishing them from real news articles. They consider an unpaired setting, where the goal if to detect whether a news article was generated by a `human` or by a `machine`, and a paired setting, where the model is given two news articles with the same meta data, one real and one machine-generated, and the detector has to assign the machine-generated article a higher `machine` probability. Thus, they essentially addressed the problem of detecting machine-generated vs. human-written news articles, even though they talked about detecting neural fake news. Later work (Pagnoni et al., 2022) discussed different threat scenarios from neural fake news generated by state-of-the-art language models and assessed the performance of generated-text detection systems under these threat scenarios. Other work proposed more advanced fake news generators that incorporated the use of propaganda techniques as part of the process (Huang et al., 2023).

With the recent popularity of LLMs, many worry about malicious actors using more powerful models such as ChatGPT, GPT3 and GPT3.5 as potential sources of machine-generated fake news and mis/dis-information(Zhou et al., 2023; Hanley and Durumeric, 2023; Su et al., 2023b).

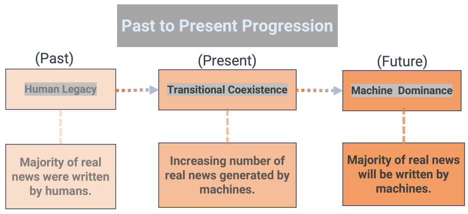


Figure 1: Our three experimental phases: (*Human Legacy*, *Transitional Coexists*, and *Machine Dominance*) based on real news generation sources.

3 Methodology

As the dynamics between human-written and machine-generated content shift, it is crucial to gauge their impact on a model's proficiency in differentiating between real and fake news. Here, we consider three distinct experimental setups, each representing different phases for news article generation due to the evolution of LLMs, as elucidated in Figure 1.

The initial *Human Legacy* stage, is emblematic of a time when the news was predominantly crafted by human authors. In this experimental setting, we used solely human-written real news articles for the training data in the real news category. Meanwhile, to see how the proportion of machine-generated fake news in the training data affects the performance of the detector, we incrementally introduce machine-generated articles into the fake news category, ranging from 0% to 100%. This setting mirrors a past era, where humans were the primary producers of real news, with machines playing a negligible role for fake news article generation.

Transitioning to the *Transitional Coexistence* stage, we reflect the current situation where language models collaboratively contribute to real news article generation. To simplify this setting, our training data in real news class contain a human-written and a machine-generated parts. This setting reflects the ongoing transformation in the news landscape, marked by the growing influence of LLMs.

Finally, in the *Machine Dominance* stage,+

4 Experiments

In this section, we introduce the dataset, the baselines, the experimental details, and the evaluation measure we use.

4.1 Datasets

We use GossipCop++ and PolitiFact++, which were introduced in (Su et al., 2023a). Table 3 shows statistics about them. The human-written fake news (**HF**) and human-written real news (**HR**) parts of the dataset are originally from the FakeNewsNet (Shu et al., 2020), and they were filtered to keep only the subset that contains a title and a description. The machine-paraphrased real news (**MR**) and the machine-generated fake news (**MF**) parts are generated by ChatGPT using *Structured Mimicry Prompting* (SMP) (Su et al., 2023a) to reduce the identifiable structure of machine-generated news articles, so that the detector can focus on the truthfulness of the content rather than on the source. More analysis and description of the dataset can be found in Appendix C.

4.2 Baselines

In our experiments, we use transformer-based methods, as they have demonstrated significantly superior performance compared to other deep learning classifiers and have gained widespread acceptance and adoption in the field of fake news detection (Kula et al., 2021a; Kong et al., 2020; Kula et al., 2021b; Kozik et al., 2023; Gundapu and Mamidi, 2021). In particular, we experimented with both large and base models of BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), ALBERT (Lan et al., 2020), and DeBERTa (He et al., 2021).

4.3 Experimental Details

We split the dataset GossipCop++ into 0.6:0.2:0.2, for training, validation, and testing, respectively. For fair evaluation and to better observe the trends, we didn't use the full dataset, but made the training/validation/testing data fully balanced by first sampling 4084 data for fake news class and 4084 data in real news class and then make the 0.6:0.2:0.2 split on them. For out-of-domain testing on PolitiFact++ dataset, we sample 97 data for each subclass for testing (i.e., 194 for real and fake news, respectively). The number of samples used in our experiments are summarized in Table 1. All models are trained on an A100 40G GPU with a batch size of 25 with a learning rate of 1e-6 for 10 epochs.

	Train		Val		Test	
Dataset	Fake	Real	Fake	Real	Fake	Real
GossipCop++	2450	2450	817	817	817	817
PolitiFact++	-	-	-	-	194	194

Table 1: Number of news articles used in our experiments.

4.4 Evaluation Measure

Since we have a balanced training and testing dataset in all the experiments, we use subclass-wise accuracy as our primary evaluation measure. Other measures such as F1, precision, recall and overall accuracy can be directly derived from the subclass-wise accuracy due to the balanced (sub)class setting. For our purposes, subclass-wise accuracy offers a more direct and insightful perspective, allowing us to assess the results from the standpoint of each individual subclass while considering more measures such as the internal bias of the detector.

5 Experimental Results

In this section, we undertake exhaustive experiments and exploration of the three stages mentioned in Section 3. Specifically, we evaluate five transformer-based models in two distinct sizes across the three stages. Coupled with the five different proportions of machine-generated fake news, this resulted in a total of 50 unique model configurations. We tested each of these configurations on two datasets: an in-domain dataset GossipCop++ and an out-of-domain dataset PolitiFact++. (As we analyzed in Appendix C, given the significant statistical differences from GossipCop++, PolitiFact++ can serve as a valuable out-of-domain dataset for assessing the robustness of the detector.)

5.1 Main Results

Given the sheer volume of the experiments, to maintain clarity and to avoid overwhelming the readers, we relegate the complete results to Appendix B, while focusing our analysis and discussion primarily on Figure 2, which shows the performance measures obtained from training a large-sized RoBERTa model and testing on the GossipCop++ dataset.

To provide a thorough understanding, we first delve into each stage independently, and then we perform a more holistic analysis of the observing patterns across these stages.

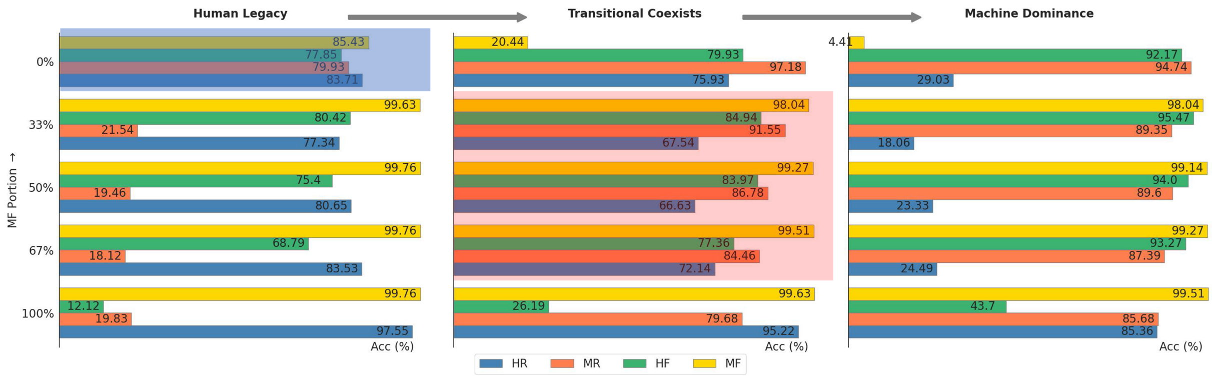


Figure 2: Class-wise detection accuracy from the *Human Legacy* stage (left), to the *Transitional Coexistence* stage (middle), to the *Machine Dominance* stage (right), with different fraction of machine generated fake news in the fake news training data illustrated in the y axis. (The blue and the red shaded area are recommended training strategies based on our experiments. We discuss this in detail in Section 6.)

***Human Legacy* Setting.** In this setting, the training data in the real news is all human-written. When paired with human-written fake news as the whole training set, it can achieve a relatively balanced and high detection accuracy for each subclass. When the MF portion increases to 33%, the detection accuracy for MF increases to around 99%, and further increases in the portion for the MF subclass in the training data almost has no more contribution to the test detection accuracy for the MF subclass. Moreover, we find an abrupt drop of detection accuracy for the MR subclass. This might be because, when we add MF to the training data, since we do not have MR data during training, the detector might use a short cut such as features that are unique to machine-generated text as features for "fake news," and thus could classify most of the MR examples as fake news. Similarly, when the fraction of MF examples increases from 67% to 100%, (i.e., we only use machine-generated fake news paired with only human-written real news as training data), we observe an abrupt drop in HF accuracy: the detectors trained in this way categorize most of the human-written fake news as real, since it checks whether the text is machine-generated as a key feature for detecting fake news. Note that, even with high MF portion, the accuracy for the MR subclass is still greater than the $1 - \text{Acc}(\textbf{MF})$, which suggests that the detector can still learn some features to identify the truthfulness of the machine generated texts rather than solely using machine-generated texts features. Otherwise, we would have $\text{Acc}(\textbf{MR}) \approx 1 - \text{Acc}(\textbf{MF})$.

One key observation from this stage is, when the proportion of MF is 0%, which corresponds to a setting where we train a detector on human-written real and fake news articles and we then deploy it to detect machine-generated real and fake news. Interestingly, the resulting detector can generalize well to distinguishing between real and fake machine-generated news, with a detection accuracy almost comparable to detecting human-written ones. This suggests that maybe it is not essential to train on machine-generated real and fake news to be able to detect them. It would certainly be helpful for the overall detection accuracy if our training data distribution aligned well with the testing data; however, in real world deployment, due to the distribution shift or due to our ignorance about the distribution of new data in a real-world scenario (for example, we do not know, how many of the news articles are machine-generated, and more importantly, this distribution might change over time due to model updates and other factors (Omar et al., 2022)), the most effective way to train the detector is to train on human-written real and fake news articles.

***Transitional Coexistence* Setting.** In this setting, the training data for the real news class is composed equally of machine-generated and human-written articles. Notably, we observe that when the fake news training data is exclusively human-written, the subclass-wise accuracy for the MF subclass is relatively low, with just 20.44% while the HF class is accurately detected with 79.93% detection accuracy. Conversely, when the fake news class is entirely MF, the accuracy for the HF subclass diminishes to a mere 26.19% while the MF accuracy is high. Echoing our prior analysis from the *Human Legacy* stage, this may be attributed to the detectors leveraging features that are indicative of an article's source (machine or human) rather than of its veracity. In the absence of HF in the training data, the detector may use a short cut and assume that all fake news are machine-generated, which results in reduced accuracy for the HF subclass. A similar situation arises when no MF data is present during training, potentially leading the detector to misclassify MF articles as real news at test time.

Moreover, even with a balanced fake news class containing half MF and half HF, the detection accuracy for the MF subclass consistently surpasses other subclasses while the accuracy for HR is the lowest. This detection accuracy is not as balanced as training on only HF and HR (see the result for *Human Legacy* Stage when the MF portion is 0%, the blue shaded area). This highlights a key insight: striving for perfect balance within each subclass during training might not yield results as good as training solely on human-generated real and fake news. However, since training with the other three subclasses (HR, HF, MF) yields better result than training with purely human-written real and fake news, the overall performance might be better (depends on the subclass distribution in the test set).

***Machine Dominance* Setting** In this setting, the entire training data for the real news class comprises MR, with no exposure to HR examples during training. When the fake news class has only HF as training examples (i.e., 0% MF portion), the detector excels in discerning HF and MR, seemingly by identifying the origin (machine or human) of the article rather than modeling its factuality. Given that modeling factuality is inherently more challenging than pinpointing the article's source, this approach compromises the detection accuracy for the MF and the HR subclasses. Remarkably, introducing a modest 33% of MF articles to the training data triggers a dramatic surge in MF detection accuracy, catapulting it from a mere 4.41% to an impressive 98.04%. This swift adaptation suggests, in this training set, that the detector has the capability to discern genuine from counterfeit content without being misled by superficial features classifying MF and MR categories. Such behavior hints at the possibility that the veracity of machine-generated articles (MF and MR) is more discernible than that of human-generated articles (HF and HR). This hypothesis can be further illuminated by comparing between the *Machine Dominance* setting (with 100% MF) and the *Human Legacy* setting (with 0% MF), where the experiments show that, detectors trained exclusively on human-written articles exhibit commendable accuracy even with machine-generated content, while, in contrast, those trained entirely on machine-generated articles often mistakenly classify the HF subclass as real.

5.2 Class-wise Accuracy as a Function of the Proportion of MF Examples

In this section, we delve into the subclass-wise accuracy for each category. Our primary focus is on understanding how accuracy trends evolve with as the proportion of MF examples increases and discerning the variations in these trends across the different stages. This analysis is visually represented in Figure 3.

The Impact of Increasing the Proportion of MF Examples We can observe in Figure 3 some consistent trends across all three stages: as the MF portion increases, the accuracy for MF and HR subclasses increases, whereas the accuracy for the HF and the MR subclasses decreases. The improvement for MF and the decrease for HF are to be expected given that the detectors are exposed to a larger number of MF examples and fewer HF examples during training. The intriguing aspect is the dip in MR detection accuracy and the boost in HR accuracy as the MF portion grows. Our hypothesis is that, when exposed with more MF training examples, the model increasingly relies on source-related features. Since MR shares confounding features with MF (because they are both machine-generated), their representations are more alike. This similarity might cause the MR examples to be misclassified more frequently as the proportion of MF examples increases. Conversely, the HR subclass, which has the least resemblance to the MF subclass, might experience improved accuracy due to the increased presence of MF examples in the training data.

Class-Wise Accuracy Across Stages. When examining subclass-wise detection rates across stages, the *Transitional Coexistence* setting consistently occupies a median position between the other two stages. Specifically, the *Machine Dominance* setting excels in detecting the HF and the MR subclasses, yet it struggles with the HR and the MF subclasses. In contrast, the *Human Legacy* setting demonstrates the prowess in accurately identifying the HR and the MF subclasses, but exhibits diminished accuracy for the HF and the MR subclasses.

Since the *Machine Dominance* setting predominantly sees machine-generated real news articles during training, it might become biased towards identifying such patterns, leading to a higher detection rate for HF and MR but lower for HR and MF. Also, if machine-generated articles have certain consistent patterns, the detector trained predominantly on MR data might rely heavily on these patterns for classification, which affects its performance on HR, which might lack these specific patterns. A similar analysis holds for the *Human Legacy* setting.

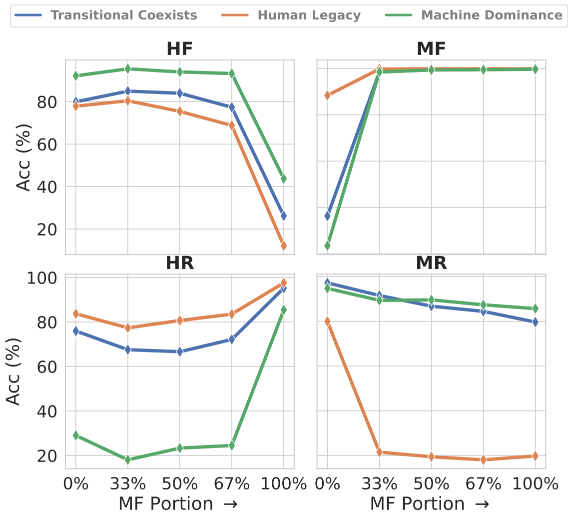


Figure 3: Illustration of the subclass-wise detection accuracy as a function of the proportion of MF examples (during training) in the three chronological settings.

5.3 Analysis of Different Detectors

Below, we compare different detectors.

Different Model Architecture. In Figure 4, we compare five detectors: fine-tuned on RoBERTa, BERT, ELECTRA, ALBERT, and DeBERTa (all large-sized models) in the *Human Legacy* setting. We can observe that no model can achieve high detection accuracy for all four subclasses. Instead, there is a trade off: a detector fine-tuned on RoBERTa achieves the highest detection accuracy in HF and MF, but the lowest accuracy for HR and MR. Meanwhile, a detector fine-tuned on ALBERT achieves the lowest detection accuracy for HF and MF, but the highest accuracy on HR and MR. Similar observations can be made about the *Transitional Coexists* and the *Machine Dominance* settings (see Appendix 11). This might be due to internal model biases: adetector fine-tuned on RoBERTa is more likely to classify an articles as fake, while such fine-tuned on ALBERT is more likely to classify it as real.

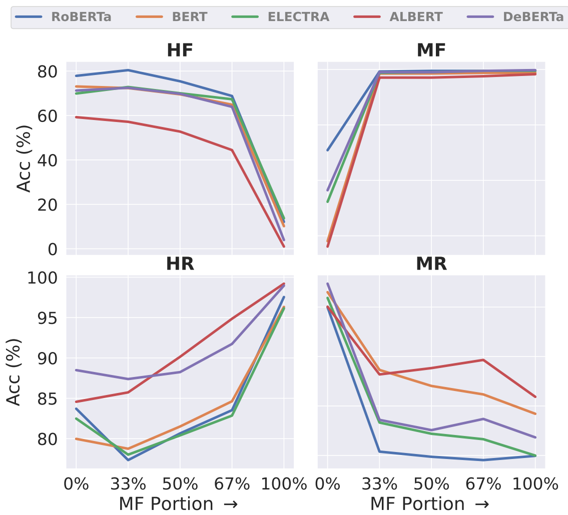


Figure 4: Comparing different detectors (RoBERTa, BERT, ELECTRA, ALBERT, DeBERTa) in the *Human Legacy* setting.

Impact of Model Size To assess how the model size affects detection outcomes, we tested both the large-sized and the base-sized versions of ALBERT and RoBERTa, as shown in Figure 5. Interestingly, a larger model does not always outperform the smaller one. In some cases, the smaller model might even mitigate the biases present in the larger variant, yielding better detection results for certain subclasses. For example, detectors trained on the large-sized ALBERT version show diminished accuracy for the HF subclass compared to the base-sized version. This disparity is even more evident for RoBERTa. Although its larger version adeptly detects HF and MF subclasses, it falters with HR and MR. Conversely, the base-sized RoBERTa model overcomes some of these biases, improving the results for HR and MR, but sacrificing the performance for HF and MF. Similar trends can also be seen in Figure 12 in the Appendix for the other stages. In summary, no single model size is universally superior. While a larger model might enhance the accuracy for certain subclasses, it might do so at the expense of other subclasses.

5.4 Out-of-Domain Detection

In this section, we evaluate the fake news detector on out-of-domain data. As shown in Figure 6, the detection accuracy has largely declined for almost all subclasses except for MR, where better or equal detection accuracy is achieved when testing on the

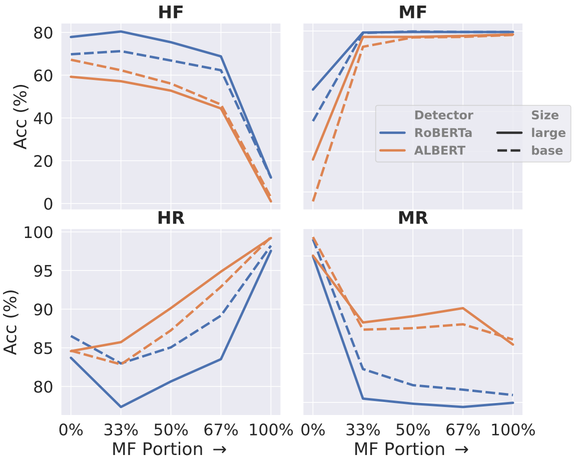


Figure 5: Comparing RoBERTa and ALBERT in the *Human Legacy* setting with large-sized and base-sized model.

out-of-domain `PolitiFact++` dataset. Also, we notice that increasing the proportion of MF examples can help mitigate the gap of out-of-domain detection accuracy at the expense of the detection accuracy for HF and MR.

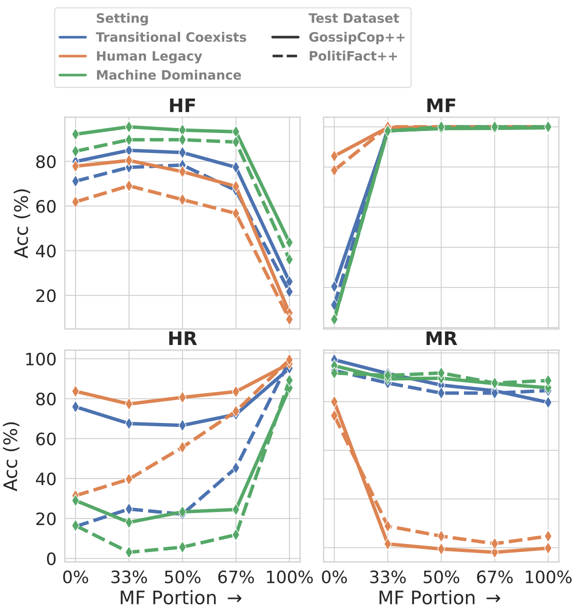


Figure 6: In-domain (`GossipCop++`) vs. out-of-domain (`PolitiFact++`) detection.

6 Discussion

The above experiments provide us with several valuable insights, which will be discussed and summarized in this section. Here, we offer some suggestions about the training data, i.e., how we should balance the machine-generated training data (MF,

Subgroup	Training Data	RoBERTa	BERT	ELECTRA	ALBERT	DeBERTa
MR	All human	-5.7	-1.51	-3.31	-3.88	-1.84
	Mixed	-3.28	-1.09	0.58	-2.89	2.9
MF	All human	-7.08	-8.21	-13.25	8.23	-21.51
	Mixed	0.73	0.21	1.35	1.33	-0.1
HR	All human	-52.27	-39.77	-7.23	-4.67	-30.24
	Mixed	-44.46	-39.17	-18.43	-0.04	-33.68
HF	All human	-15.99	-18.43	-22.47	-6.66	-16.6
	Mixed	-5.62	-11.33	-11.85	-23.51	-4.75

Table 2: Performance degradation in out-of-domain compared to in-domain detection when training on all human data and on mixed data in proportion of HF:MF:HR:MR=1:1:1:1.

MR) and the human training data (HF, HR) when training the detector.

6.1 In-Domain Detection

First, we found that training with either all human written data (see the left most subfigure of Figure 2 where we highlighted with blue shades) or with a mixture of all four subclasses (see the middle subfigure in Figure 2, which are highlighted with red shades) can achieves a relatively satisfying detection result on all the subclasses. However, detectors trained with all human written data (the blue shaded part) seem to be a better option since it is more balanced on each subclass while detectors trained on some mixtures of all four subclasses (the red shaded area) scarifies HR accuracy for higher MF detection accuracy). Thus, we recommend using only human real and fake new articles for training in domain detector.

6.2 Out-of-Domain Detection

As indicated in Figure 6, when increasing MF portion, the distance of in-domain detection accuracy and out-of-domain accuracy becomes smaller. To verify this quantitatively, we calculated the gap of in-domain detection accuracy and out-of-domain accuracy (namely, the class-wise accuracy of `PolitiFact++` minus the class-wise accuracy of `GossipCop++`), when trained with only human written news articles as well as trained with mixed sources (HF:MF:HR:MR=1:1:1:1). The results are illustrated in Table 2, where we found that, using mixed training data sources leads to a smaller gap of in-domain and out-of-domain detection accuracy. Thus, it is suggested to train a detector by adding some portion of machine generated real and fake news data to improve the detectors' generalization ability on different domains.

7 Limitations

One limitation of our study is that we studied a coarse-grained proportion of machine-generated articles in the training data. Our primary objective was to offer insights and to highlight potential adaptations in the training strategies during the LLM era, thus raising awareness of responsible use of LLMs, and the three stages we outlined. Note that it is easy to extend our framework to a more fine-grained study.

The limitation in our paper as well as the observation from the experiments evoke several interesting future directions to address. From the perspective of fake news detection and misinformation research, there is a need for more nuanced evaluation and for combining different detectors to improve the detection accuracy for better fake news detection. Moreover, our experiments inspire us to generalize the study of real/fake news distribution drift trends to macro contexts, particularly in light of how LLMs influence data distribution shifts. We elaborate more on this below.

More Fine-Grained Evaluation Setting. Our experiments revealed that while training exclusively on human-generated data yields balanced and high accuracy for each subclass relative to the mixed training approach, its robustness is limited for out-of-domain detection. Incorporating some machine-generated data appears to enhance this robustness without significant performance trade-offs. Our current study focused only on the MR proportions of 0%, 50%, and 100%. Further, nuanced experiments are required to pinpoint the optimal balance between class-specific detection accuracy and robustness. It is particularly pertinent to explore MR proportions under 50% to assess performance and robustness.

Combining Different Detectors. As detailed in Subsection 5.3, different detectors exhibit different level of biases towards the individual subclasses. Leveraging ensembling to amalgamate these detectors could offset some inherent biases, potentially leading to enhanced accuracy across the classes.

Data Distribution Shift and its Consequences. Our paper delineates three temporal settings: *Human Legacy*, *Transitional Coexistence*, and *Machine Dominance*. These stages offer a simplified view of potential LLM-induced distribution changes, when observed in a longer time span. One angle to approach this data distribution shift is via performative prediction (Perdomo et al., 2020), suggesting that model outputs reciprocally influence data distribution. While there is still a discernible gap remains between human-written and machine-generated text distributions, the pervasive use of LLMs and their outputs might influence the human text distribution, and over time, the relative proportion of machine-generated and human-written texts would get closer to each other and might converge to a static landscape. For example, in Figure 9, we can observe a distinctive discrepancy with MR and MF, while HF and HR are quite similar. We conjecture that the distribution of the four subclasses might evolve to convergence given a sufficient time horizon. Thus, it woudl be interesting to analyze fake news detection within this evolving framework.

8 Ethics and Broader Impact

Our research delves into fake news detectors and the dynamics of mis/disinformation, positing three hypothetical scenarios. While these settings are grounded in reason, they primarily serve to gauge detector performance and behavior. They should not be construed as predictions of the future landscape of fake and real news generation. Our aim is to raise awareness of the potential risks that LLMs can induce, which goes beyond mis/disinformation and fake news detection, but to more subtle ways of influence related to the proportion of human-written texts online. We advocate for a responsible use of LLMs.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *ArXiv preprint*, abs/2108.07258.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33:*

Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, lies, and automation. *Center for Security and Emerging Technology*, 1(1):2.

Sonia Castelo, Thais Almeida, Anas Elghafari, Aécio Santos, Kien Pham, Eduardo Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 World Wide Web conference*, pages 975–980.

Yimin Chen, Nadia K Conroy, and Victoria L Rubin. 2015. News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *ArXiv preprint*, abs/2301.04246.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Lucas Graves and Federica Cherubini. 2016. The rise of fact-checking sites in europe. *Digital News Project Report*.

Sunil Gundapu and Radhika Mamidi. 2021. Transformer based automatic covid-19 fake news detection system. *ArXiv preprint*, abs/2101.00180.

Hans WA Hanley and Zakir Durumeric. 2023. Machine-made media: Monitoring the mobilization of machine-generated articles on misinformation and mainstream news websites. *ArXiv preprint*, abs/2305.09820.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pages 324–332. ACM.

Sheng How Kong, Li Mei Tan, Keng Hoon Gan, and Nur Hana Samsudin. 2020. Fake news detection using deep learning. In *2020 IEEE 10th symposium on computer applications & industrial electronics (ISCAIE)*, pages 102–107. IEEE.

Rafał Kozik, Aleksandra Pawlicka, Marek Pawlicki, Michał Choraś, Wojciech Mazurczyk, and Krzysztof Cabaj. 2023. A meta-analysis of state-of-the-art automated fake news detection methods. *IEEE Transactions on Computational Social Systems*.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Sebastian Kula, Michał Choraś, and Rafał Kozik. 2021a. Application of the bert-based architecture in fake news detection. In *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12*, pages 239–249. Springer.

Sebastian Kula, Rafał Kozik, Michał Choraś, and Michał Woźniak. 2021b. Transformer based models in fake news detection. In *International Conference on Computational Science*, pages 28–38. Springer.

Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *ArXiv preprint*, abs/1804.08559.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4826–4832. ijcai.org.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *ArXiv preprint*, abs/2301.11305.

Marwan Omar, Soohyeon Choi, DaeHun Nyang, and David Mohaisen. 2022. Quantifying the performance of adversarial training on language models with distribution shifts. In *Proceedings of the 1st Workshop on Cybersecurity and Social Sciences*, pages 3–9.

Ray Oshikawa, Jing Qian, and William Yang Wang. 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.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Demetris Paschalides, Alexandros Kornilakis, Chrysovalantis Christodoulou, Rafael Andreou, George Pallis, Marios Dikaiakos, and Evangelos Markatos. 2019. Check-it: A plugin for detecting and reducing the spread of fake news and misinformation on the web. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 298–302.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 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.

Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A platform for tracking online misinformation. In *Proceedings of the 25th international conference companion on world wide web*, pages 745–750.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis) informs us better than humans. *ArXiv preprint*, abs/2301.11924.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023a. Fake news detectors are biased against texts generated by large language models. *ArXiv preprint*, abs/2309.08674.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023b. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. *ArXiv preprint*, abs/2306.05540.

Chris J Vargo, Lei Guo, and Michelle A Amazeen. 2018. The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New media & society*, 20(5):2028–2049.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Zhou X Jain A Phoha VV and R Zafarani. 2020. Fake news early detection: a theory-driven model. *Digital Threats: Research and Practice*, 1(2):1.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9051–9062.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *ArXiv preprint*, abs/2303.18223.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

A Original Dataset statistics

Dataset	HF	MF	HR	MR
`GossipCop++`	4,084	4,084	8,168	4,169
`PolitiFact++`	97	97	194	132

Table 3: Number of news articles from each subclass in the `GossipCop++` and `PolitiFact++` datasets.

B Complete Results

The complete results for the three stages evaluated in our paper are shown in the tabels below: for the *Human Legacy* setting in Table 4, for the *Transitional Coexists* setting in Table 5, and for the *Machine dominance* setting in Table 6. We show results when using different detectors for in-domain (`GossipCop++`) and out-of-domain (`PolitiFact++`) experiments.

MF portion (Training Data)	Model size	Model name	`GossipCop++` Accurancy w.r.t. each group Real HR	MR	Fake HF	MF	`PolitiFact++` Accurancy w.r.t. each group Real HR	MR	Fake HF	MF
0%	Large	RoBERTa	83.71	79.93	**77.85**	**85.43**	31.44	74.23	61.86	78.35
		BERT	79.98	86.05	73.07	69.03	40.21	84.54	54.64	60.82
		ELECTRA	82.49	83.72	69.89	76.13	75.26	80.41	47.42	62.89
		ALBERT	84.57	80.17	59.24	68.05	79.90	76.29	52.58	76.29
		DeBERTa	**88.49**	**89.47**	71.24	78.21	58.25	87.63	54.64	56.70
	Base	RoBERTa	86.53	86.90	69.77	77.60	77.84	84.54	37.11	61.86
		BERT	86.28	84.33	63.16	78.70	76.80	85.57	30.93	69.07
		ELECTRA	86.83	82.86	63.53	80.66	**90.72**	80.41	40.21	**79.38**
		ALBERT	84.63	87.76	67.20	57.65	65.46	**88.66**	57.73	56.70
		DeBERTa	80.47	81.52	70.13	78.09	70.10	79.38	**74.23**	78.35
33%	Large	RoBERTa	77.34	21.54	**80.42**	**99.63**	39.69	28.87	**69.07**	**100.00**
		BERT	78.75	**54.59**	72.34	99.27	44.33	50.52	60.82	97.94
		ELECTRA	78.02	33.29	72.83	99.39	72.68	31.96	59.79	98.97
		ALBERT	85.73	52.75	57.16	98.53	81.96	**51.55**	31.96	97.94
		DeBERTa	**87.39**	34.39	72.46	99.51	72.16	42.27	64.95	**100.00**
	Base	RoBERTa	82.98	33.66	71.24	99.51	73.71	25.77	50.52	**100.00**
		BERT	83.71	46.14	65.97	99.39	64.95	47.42	36.08	**100.00**
		ELECTRA	83.28	37.33	63.04	97.92	**89.69**	35.05	48.45	**100.00**
		ALBERT	82.85	49.82	62.30	96.08	71.13	50.52	40.21	97.94
		DeBERTa	87.08	39.29	64.63	98.65	81.96	36.08	62.89	98.97
50%	Large	RoBERTa	80.65	19.46	**75.40**	99.76	55.67	24.74	**62.89**	**100.00**
		BERT	81.51	48.10	69.52	99.27	45.88	46.39	51.55	97.94
		ELECTRA	80.40	28.76	70.01	99.51	82.99	27.84	52.58	**100.00**
		ALBERT	**90.14**	**55.32**	52.75	98.53	**91.75**	**53.61**	27.84	98.97
		DeBERTa	88.24	30.23	69.77	99.51	64.95	34.02	57.73	**100.00**
	Base	RoBERTa	85.06	27.05	66.83	**99.88**	83.51	23.71	40.21	**100.00**
		BERT	85.73	44.68	62.67	99.39	70.10	46.39	34.02	**100.00**
		ELECTRA	85.55	33.41	61.32	99.27	91.24	30.93	42.27	**100.00**
		ALBERT	87.26	50.43	56.06	98.41	81.96	51.55	31.96	**100.00**
		DeBERTa	89.83	35.74	59.61	99.27	90.21	32.99	47.42	**100.00**
67%	Large	RoBERTa	83.53	18.12	**68.79**	**99.76**	73.71	21.65	**56.70**	**100.00**
		BERT	84.63	44.68	64.87	99.39	60.31	39.18	40.21	97.94
		ELECTRA	82.85	26.56	67.32	99.76	88.66	26.80	45.36	**100.00**
		ALBERT	**94.86**	**58.63**	44.43	98.78	96.91	**59.79**	20.62	98.97
		DeBERTa	91.73	34.76	63.89	99.76	75.26	38.14	47.42	**100.00**
	Base	RoBERTa	89.16	25.21	62.30	99.76	90.21	23.71	29.90	**100.00**
		BERT	87.75	44.31	55.20	99.51	78.35	45.36	26.80	**100.00**
		ELECTRA	88.36	34.27	57.65	99.39	94.85	32.99	30.93	**100.00**
		ALBERT	92.90	52.02	46.27	98.53	92.27	52.58	20.62	**100.00**
		DeBERTa	92.77	29.99	47.37	99.39	**97.42**	28.87	35.05	**100.00**
100%	Large	RoBERTa	97.55	19.83	12.12	99.76	99.48	24.74	9.28	**100.00**
		BERT	96.33	36.84	10.16	99.39	87.63	34.02	**12.37**	**100.00**
		ELECTRA	96.14	19.95	13.71	99.76	99.48	25.77	6.19	**100.00**
		ALBERT	99.20	43.70	0.98	99.14	98.97	**49.48**	1.03	98.97
		DeBERTa	98.96	27.29	3.92	**99.88**	99.48	34.02	9.28	**100.00**
	Base	RoBERTa	98.22	23.01	12.12	99.76	98.97	25.77	3.09	**100.00**
		BERT	98.16	41.74	6.61	99.76	96.39	43.30	4.12	**100.00**
		ELECTRA	94.67	28.52	**18.97**	99.76	97.42	28.87	8.25	**100.00**
		ALBERT	**99.33**	**45.78**	2.82	99.02	**100.00**	48.45	4.12	**100.00**
		DeBERTa	98.53	28.03	7.83	99.76	**100.00**	32.99	8.25	**100.00**

Table 4: Complete result in the *Human Legacy* setting.

MF portion (Training Data)	Model size	Model name	GossipCop++				PolitiFact++			
			Accurancy w.r.t. each group				Accurancy w.r.t. each group			
			Real		Fake		Real		Fake	
			HR	MR	HF	MF	HR	MR	HF	MF
0%	Large	RoBERTa	75.93	97.18	**79.93**	20.44	15.98	92.78	71.13	11.34
		BERT	78.08	97.43	74.30	14.32	36.60	**97.94**	60.82	**15.46**
		ELECTRA	81.38	97.31	72.34	**27.29**	30.93	94.85	68.04	6.19
		ALBERT	65.52	92.53	73.68	13.34	51.55	90.72	73.20	15.46
		DeBERTa	75.81	96.33	77.23	24.72	39.69	91.75	61.86	4.12
	Base	RoBERTa	79.79	97.67	73.19	25.34	68.04	96.91	51.55	13.40
		BERT	78.02	96.94	68.67	18.85	65.98	95.88	59.79	7.22
		ELECTRA	**84.75**	**98.04**	66.10	19.09	**84.54**	95.88	46.39	1.03
		ALBERT	66.69	94.61	74.66	17.01	36.60	93.81	73.20	9.28
		DeBERTa	63.99	94.61	79.07	18.36	40.72	89.69	**78.35**	7.22
33%	Large	RoBERTa	67.54	91.55	**84.94**	**98.04**	24.74	87.63	77.32	**98.97**
		BERT	62.46	86.66	82.99	95.35	18.04	84.54	72.16	95.88
		ELECTRA	70.73	91.19	79.19	96.33	40.72	87.63	68.04	97.94
		ALBERT	69.38	89.84	68.05	91.06	66.49	84.54	53.61	91.75
		DeBERTa	69.63	93.76	80.29	97.06	47.42	**92.78**	**81.44**	95.88
	Base	RoBERTa	70.12	89.84	79.93	93.15	50.52	89.69	56.70	88.66
		BERT	74.59	92.04	74.05	95.47	41.75	91.75	63.92	98.97
		ELECTRA	72.99	89.84	72.58	88.37	**78.87**	87.63	68.04	91.75
		ALBERT	72.32	92.53	72.46	89.60	44.33	90.72	72.16	95.88
		DeBERTa	**74.83**	**94.12**	73.68	91.19	48.97	87.63	80.41	88.66
50%	Large	RoBERTa	66.63	86.78	**83.97**	**99.27**	22.16	83.51	**78.35**	**100.00**
		BERT	71.65	86.66	78.34	96.70	32.47	85.57	67.01	96.91
		ELECTRA	71.52	89.11	75.76	98.65	53.09	89.69	63.92	**100.00**
		ALBERT	**79.42**	91.55	57.53	93.51	79.38	88.66	34.02	94.85
		DeBERTa	76.97	**94.00**	75.89	98.04	43.30	**96.91**	71.13	97.94
	Base	RoBERTa	74.89	88.13	77.23	95.84	55.67	83.51	54.64	92.78
		BERT	78.44	90.82	70.50	96.82	54.64	91.75	55.67	98.97
		ELECTRA	77.83	87.39	67.32	93.88	**85.57**	90.72	58.76	94.85
		ALBERT	78.81	91.06	64.38	91.92	68.04	88.66	45.36	95.88
		DeBERTa	76.67	92.41	70.13	94.74	66.49	85.57	77.32	94.85
67%	Large	RoBERTa	72.14	84.46	**77.36**	**99.51**	45.36	83.51	**67.01**	**100.00**
		BERT	76.06	84.70	72.71	98.65	39.18	83.51	60.82	97.94
		ELECTRA	74.65	88.74	71.60	99.39	77.32	89.69	53.61	**100.00**
		ALBERT	**87.32**	92.41	45.90	95.47	88.66	92.78	17.53	94.85
		DeBERTa	84.63	**95.10**	65.97	99.14	77.32	**94.85**	58.76	**100.00**
	Base	RoBERTa	76.55	84.82	73.56	98.90	75.26	82.47	40.21	98.97
		BERT	84.38	90.21	63.16	97.80	72.68	90.72	37.11	98.97
		ELECTRA	81.14	86.78	62.30	96.45	88.14	88.66	46.39	98.97
		ALBERT	86.65	92.17	54.10	95.10	80.93	91.75	35.05	94.85
		DeBERTa	85.06	89.23	53.12	95.96	**92.27**	88.66	44.33	97.94
100%	Large	RoBERTa	95.22	79.68	**26.19**	**99.63**	98.97	84.54	**21.65**	**100.00**
		BERT	96.02	83.48	14.81	98.41	84.02	80.41	17.53	98.97
		ELECTRA	95.71	86.17	21.54	99.63	96.91	84.54	16.49	**100.00**
		ALBERT	**99.27**	**96.08**	1.96	96.57	**99.48**	**97.94**	2.06	95.88
		DeBERTa	98.53	93.88	9.18	99.39	99.48	93.81	18.56	**100.00**
	Base	RoBERTa	95.41	78.09	24.24	99.63	97.42	76.29	6.19	**100.00**
		BERT	96.39	86.05	9.91	98.41	90.21	85.57	11.34	**100.00**
		ELECTRA	93.75	85.31	25.21	98.29	95.88	85.57	16.49	**100.00**
		ALBERT	98.53	95.72	5.14	96.70	97.42	96.91	3.09	96.91
		DeBERTa	97.80	92.41	11.75	98.90	98.45	92.78	12.37	98.97

Table 5: Complete result in the *Transitional Coexists* setting.

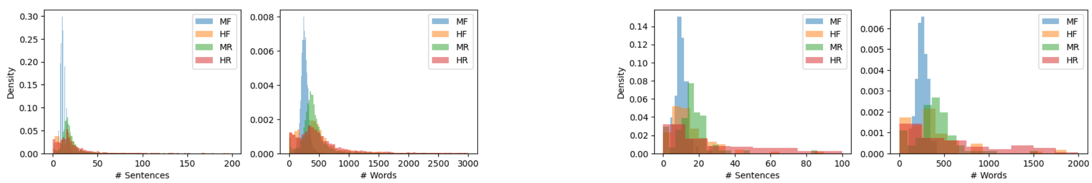
MF portion (Training Data)	Model size	Model name	GossipCop++				PolitiFact++			
			Accurancy w.r.t. each group				Accurancy w.r.t. each group			
			Real		Fake		Real		Fake	
			HR	MR	HF	MF	HR	MR	HF	MF
0%	Large	RoBERTa	29.03	94.74	92.17	4.41	16.49	91.75	84.54	4.12
		BERT	38.09	93.76	89.47	3.67	23.20	93.81	82.47	**7.22**
		ELECTRA	**39.07**	95.10	86.29	10.77	12.89	94.85	81.44	2.06
		ALBERT	16.35	87.64	**94.86**	6.98	17.53	86.60	**91.75**	6.19
		DeBERTa	24.68	**96.21**	93.27	7.96	13.92	**95.88**	90.72	3.09
	Base	RoBERTa	27.62	92.66	89.11	9.67	13.40	88.66	84.54	3.09
		BERT	29.94	91.43	85.68	6.73	**25.77**	91.75	81.44	6.19
		ELECTRA	34.05	93.15	84.94	3.79	22.16	92.78	86.60	1.03
		ALBERT	19.41	90.45	93.02	7.96	16.49	89.69	90.72	4.12
		DeBERTa	17.33	91.80	94.49	**14.20**	11.34	87.63	89.69	6.19
33%	Large	RoBERTa	18.06	89.35	95.47	**98.04**	3.09	90.72	89.69	**97.94**
		BERT	22.11	86.41	94.49	95.72	10.31	79.38	89.69	97.94
		ELECTRA	**30.25**	**92.41**	91.31	89.35	9.28	**91.75**	90.72	91.75
		ALBERT	15.74	83.72	94.12	91.80	15.46	82.47	90.72	92.78
		DeBERTa	18.74	91.55	**95.72**	96.21	12.89	89.69	**96.91**	96.91
	Base	RoBERTa	26.15	89.60	92.04	92.29	**18.56**	83.51	82.47	93.81
		BERT	25.66	87.27	91.31	93.15	9.28	87.63	88.66	95.88
		ELECTRA	23.03	87.76	91.31	87.03	12.89	86.60	92.78	90.72
		ALBERT	19.17	86.90	94.74	89.60	7.22	81.44	95.88	91.75
		DeBERTa	20.58	88.74	93.27	91.06	11.34	85.57	91.75	92.78
50%	Large	RoBERTa	23.33	89.60	**94.00**	**99.14**	5.67	91.75	89.69	**100.00**
		BERT	25.41	85.31	91.55	97.31	10.82	83.51	88.66	**100.00**
		ELECTRA	**32.21**	91.55	90.21	94.12	13.92	91.75	86.60	95.88
		ALBERT	20.70	85.43	90.33	93.64	23.20	83.51	86.60	95.88
		DeBERTa	27.86	**94.00**	92.41	97.67	**25.26**	**92.78**	89.69	98.97
	Base	RoBERTa	29.58	88.13	90.21	94.74	22.16	81.44	83.51	95.88
		BERT	31.72	86.41	89.23	96.08	9.28	86.60	86.60	97.94
		ELECTRA	27.80	87.15	90.58	93.51	21.65	86.60	88.66	94.85
		ALBERT	23.82	88.37	91.19	94.86	9.79	87.63	**92.78**	97.94
		DeBERTa	22.90	85.07	90.94	89.72	24.23	87.63	90.72	94.85
67%	Large	RoBERTa	24.49	87.39	**93.27**	**99.27**	11.86	87.63	88.66	**100.00**
		BERT	34.35	84.70	89.35	97.55	12.89	83.51	81.44	**100.00**
		ELECTRA	39.25	91.55	85.43	97.31	24.74	90.72	80.41	96.91
		ALBERT	30.92	85.56	83.11	95.59	39.18	84.54	75.26	95.88
		DeBERTa	30.13	**94.49**	90.70	98.78	26.29	**95.88**	**90.72**	**100.00**
	Base	RoBERTa	34.29	88.86	86.78	96.94	38.66	81.44	75.26	97.94
		BERT	**40.54**	88.00	84.82	97.18	22.16	88.66	81.44	98.97
		ELECTRA	33.19	86.41	89.11	96.33	39.18	82.47	82.47	95.88
		ALBERT	34.97	87.76	85.92	94.61	21.65	86.60	83.51	95.88
		DeBERTa	28.23	84.82	88.13	93.39	**47.94**	87.63	85.57	95.88
100%	Large	RoBERTa	85.36	85.68	43.70	99.51	89.18	88.66	36.08	**100.00**
		BERT	90.39	90.09	26.93	98.16	69.07	89.69	28.87	98.97
		ELECTRA	89.28	92.04	31.21	99.39	86.08	89.69	27.84	**100.00**
		ALBERT	**98.22**	**97.31**	5.14	95.84	**96.39**	**100.00**	3.09	92.78
		DeBERTa	91.79	93.76	23.99	99.51	83.51	92.78	**39.18**	98.97
	Base	RoBERTa	83.28	84.33	**46.88**	**99.63**	87.11	83.51	19.59	**100.00**
		BERT	91.18	90.94	18.36	97.92	86.08	92.78	21.65	98.97
		ELECTRA	84.57	89.23	39.29	97.31	84.54	89.69	34.02	**100.00**
		ALBERT	96.14	96.82	11.14	95.96	94.33	97.94	10.31	94.85
		DeBERTa	87.32	88.98	33.17	96.70	93.81	90.72	31.96	**100.00**

Table 6: Complete results in the *Machine dominance* setting.

C Detailed Dataset Analysis

In Figure 8, we illustrate the average sentence count and word count for both GossipCop++ and PolitiFact++. We observe that HR generally consists of longer articles compared to other subclasses, while machine-generated news articles tend to be shorter on average, especially MF. Moreover, the graph demonstrates substantial variations in terms of average length across the different datasets. For instance, when comparing GossipCop++ to PolitiFact++, the former has an average of 625 words and 25 sentences, whereas the latter is significantly longer, with 3,759 words and 191 sentences, i.e., seven times larger. Another distinct difference between these two datasets is that in GossipCop++ the average sentence count and word count for HF (22 sentences and 564 words) and HR are quite close to each other. In contrast, within the PolitiFact++ dataset, HR is roughly 10 times longer than HF, with HR consisting of 17 sentences and 459 words. Although the total number of news articles in PolitiFact++ is too small to train a reliable fake news detector, it serves as a valuable out-of-domain dataset for assessing the robustness of the detector, given its significant statistical differences from GossipCop++.

In Figure 7, we extract 4,084 articles in each subclass for GossipCop++ and 97 articles in each subclass of PolitiFact++ to visualize the distribution of the number of sentences and the number of words for each subclass. See also Figure 9 and Figure 10 in the appendix. From Figure 7, we find that the distribution of sentence count and the word count for HF and HR are quite close to each other, spanning a wide range of lengths. Meanwhile, the sentence count and the word count for machine-generated articles, especially MF news articles have more pronounced peaks.



(a) GossipCop++ (b) PolitiFact++

Figure 7: Sentence count and word count density histogram for GossipCop++ and PolitiFact++.

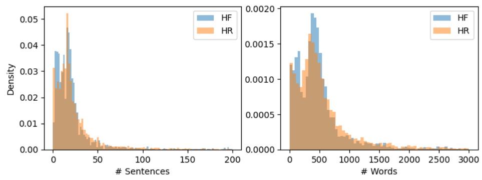
(a) GossipCop++ (b) PolitiFact++

Figure 8: Average sentence count and average word count density histogram for GossipCop++ and PolitiFact++.

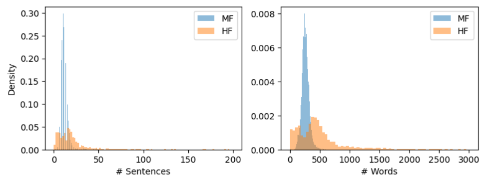
C.1 Sentence Length and Word Length

Figure 9 and Figure 10 compare the pair-wise distribution of the sentence count and the word count, from which we can observe that the distribution of sentence count and word count for HF and HR exhibit remarkable similarity. This observation implies that human-written news articles, regardless of their authenticity, share a significant resemblance in their structural composition. Conversely, there exists a more pronounced disparity in the case of machine-generated news articles (MF and MR), implying that it might be easier to distinguish the veracity of such articles based on their length distribution. Moreover, we observed a notable discrepancy in the distribution of MR and HR, even though MR is paraphrased from real news articles with an approximately the same sentence and word counts.

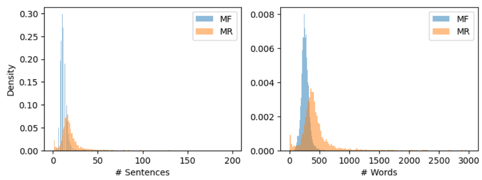
Although the dataset statistics show the distribution discrepancy between human-written and machine-generated real and fake news, which might be a signal for current fake news detection problem, from a broader data distribution standpoint, if journalists increasingly adopt LLMs in their writing, over time, the distribution of real news articles might gradually shift towards the distribution of the machine-generated articles (MF and MR). Eventually, this shift could lead to a convergence where the distributions of real and fake news articles once again closely resemble each other.



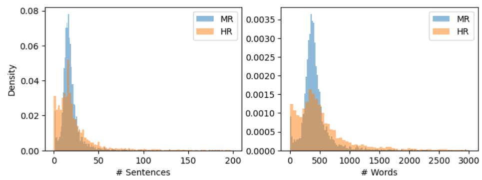
(a) HF vs. HR



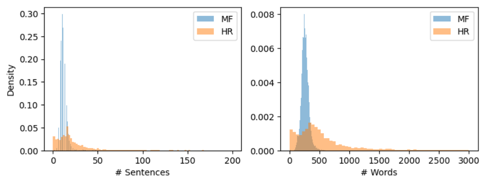
(b) MF vs. HF



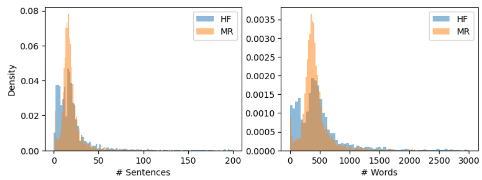
(c) MF vs. MR



(d) MR vs. HR

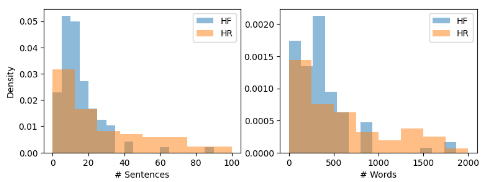


(e) MF vs. HR

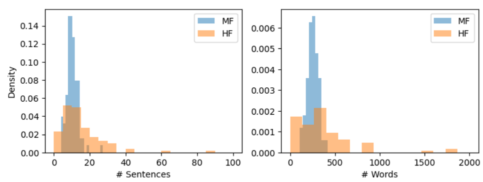


(f) HF vs. MR

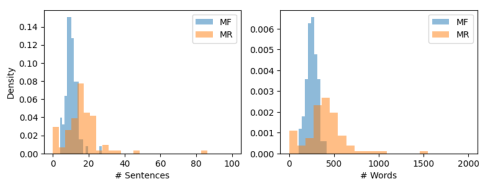
Figure 9: Comparing the sentence length and teh word length density histograms for different subclasses in `GossipCop++`.



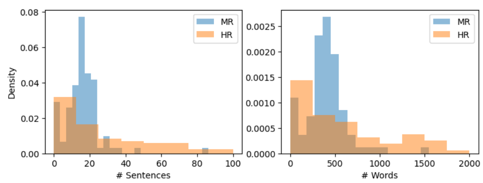
(a) HF vs. HR



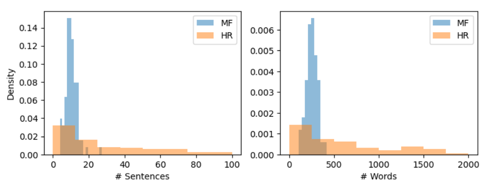
(b) MF vs. HF



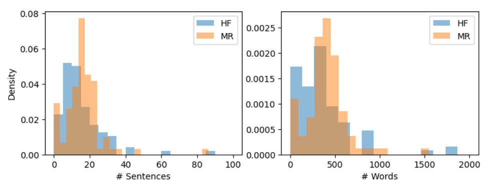
(c) MF vs. MR



(d) MR vs. HR



(e) MF vs. HR



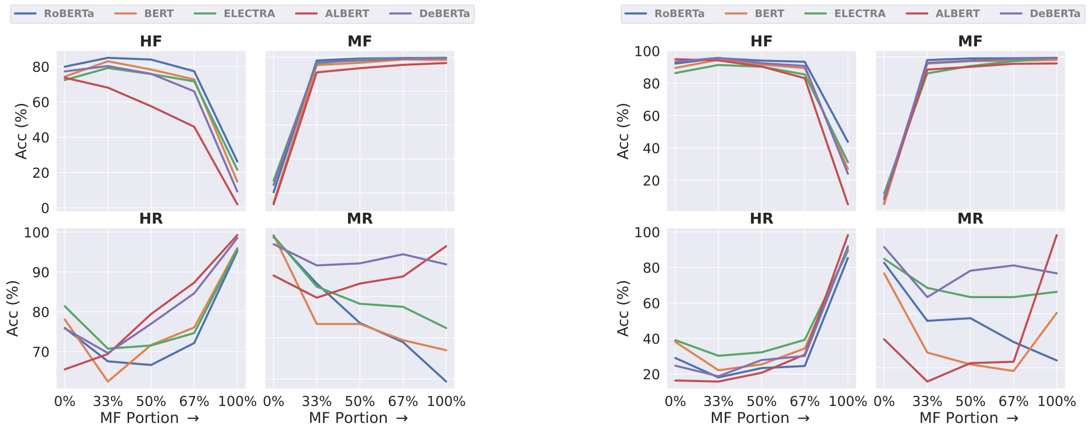
(f) HF vs. MR

Figure 10: Comparing the sentence length and the word length density histogram for different subclasses in `PolitiFact++`.

D Comparing Different Detectors in the *Transitional Coexistence* and the *Machine Dominance* Setting.

Here, we compare different detectors in the *Transitional Coexistence* and the *Machine Dominance* Setting as supplementary experiments for Section 5.3.

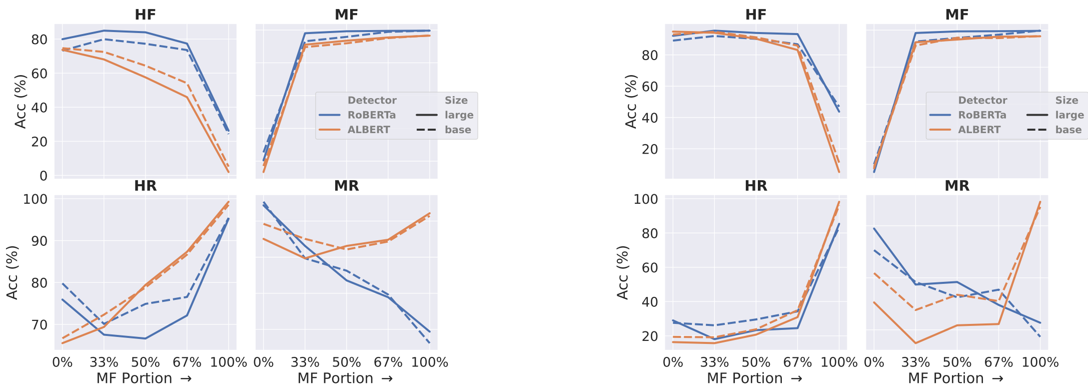
D.1 Impact of the Detector Structure



(a) *Transitional Coexistence* (b) *Machine Dominance*

Figure 11: Comparing different detectors (RoBERTa, BERT, ELECTRA, ALBERT, DeBERTa) in the *Transitional Coexists* and the *Machine Dominance* settings.

D.2 Impact of the Detector Size



(a) *Transitional Coexistence* (b) *Machine Dominance*

Figure 12: Comparing ReBERTa and ALBERT detectors in the *Transitional Coexists* and the *Machine Dominance* setting with different sizes: large and base models.