Mitigation of Diachronic Bias in Fake News Detection Dataset

Taichi Murayama and **Shoko Wakamiya** and **Eiji Aramaki**
Nara Institute of Science and Technology (NAIST)
{murayama.taichi.mk1, wakamiya, aramaki}@is.naist.jp

Abstract

Fake news causes significant damage to society. To deal with these fake news, several studies on building detection models and arranging datasets have been conducted. Most of the fake news datasets depend on a specific time period. Consequently, the detection models trained on such a dataset have difficulty detecting novel fake news generated by political changes and social changes; they may possibly result in biased output from the input, including specific person names and organizational names. We refer to this problem as **Diachronic Bias** because it is caused by the creation date of news in each dataset. In this study, we confirm the bias, especially proper nouns including person names, from the deviation of phrase appearances in each dataset. Based on these findings, we propose masking methods using Wikidata to mitigate the influence of person names and validate whether they make fake news detection models robust through experiments with in-domain and out-of-domain data.

1 Introduction

Fake news, which refers to intentional and verifiable false news stories, has caused significant damage to society. For example, Bovet and Makse (2019) noted that 25% of the news stories linked in tweets posted just before the 2016 U.S. presidential election were either fake or extraordinarily biased. In addition to elections, fake news tends to spread after unusual situations such as natural disasters (Hashimoto et al., 2020) and disease outbreaks (e.g., COVID-19 (Shahi and Nandini, 2020)). To address these problems, various studies on the development of models for fake news detection from social media posts and news content, and the construction of fake news datasets for this purpose have been conducted (Shu et al., 2020).

Most datasets for fake news detection consist of factual and fake news that actually diffuse over the Internet. The topics and contents of fake news change over time because they are strongly influenced by the interests of the general population (Schmidt et al., 2017). For instance, there were fake news related to President Obama in 2013 (CNBC, 2013), presidential election in 2016 (Bovet and Makse, 2019), and COVID-19 in 2020 (Shahi et al., 2020). Thus, datasets are frequently constructed based on fake news present in a specific period. Fake news detection models learned from these datasets achieve high accuracy for the datasets constructed for the same period and domain, while they lead to a drop in the detection performance of fake news in different domains and future applications because of the difference in word appearance. In other words, fake news detection models learned from a dataset that includes news only from a specific period may possibly result in a biased judgment from the input, based on cues related to a particular person name or an organizational name. For example, a model learned from a dataset in 2017 is difficult to correctly classify articles including “Donald Trump” or “Joe Biden” in 2021, because the model does not know a new president change. We call the problem as **Diachronic Bias** because it is caused by the difference in the news publishing date in each dataset. This problem occurs with data from even the same domain, making it difficult to construct a robust detection model.

This study examines various strategies to mitigate diachronic bias by masking proper nouns that tend to cause the bias, such as names of people and places. First, we analyzed the correlation between labels and phrases in several fake news detection datasets with different creation periods and noted the deviation of words, mainly person names. We then applied and validated several masking methods focusing on proper nouns to mitigate diachronic bias in the tackling of fake news detection tasks.

MultiFC						Constraint					
Real			Fake			Real			Fake		
Bi-gram	**LMI**	$p(l\|w)$	**Bi-gram**	**LMI**	$p(l\|w)$	**Bi-gram**	**LMI**	$p(l\|w)$	**Bi-gram**	**LMI**	$p(l\|w)$
mitt romney	218	0.69	health care	631	0.64	url url	1378	0.77	a video	591	1.0
if you	217	0.70	**barack obama**	365	0.69	rt @user	822	0.93	**donald trump**	569	0.98
rhode island	190	0.75	**president barack**	337	0.70	total number	650	0.98	has been	569	0.52
new jersey	177	0.67	**scott walker**	258	0.81	more than	635	0.89	**url donaldtrump**	435	1.0
john mccain	167	0.73	says president	218	0.78	have been	575	0.82	**bill gates**	355	1.0
no. 1	128	0.86	care law	185	0.80	@user url	449	0.87	video shows	346	0.98
voted against	128	0.71	will be	162	0.63	managed isolation	402	1.0	**president trump**	315	1.0
any other	125	0.61	**hillary clinton**	159	0.67	our daily	385	0.99	covid vaccine	293	0.80
does not	119	0.71	**gov. scott**	148	0.72	states reported	373	1.0	corona virus	275	1.0
this year	116	0.75	social security	144	0.68	update published	367	1.0	social media	275	0.93

Table 1: Top 10 LMI ranked bi-grams in MultiFC and Constraint for real and fake labels with their $p(l|w)$. **LMI** are written as value multiplied by 10^6. Person names are written in bold. There is a tendency for real labels to be highly correlated with common phrases, while fake labels are highly correlated with person names.

2 Related Work

Analysis and examination of mitigation methods for various types of bias have been conducted for detecting offensive language and hate speech: author bias (Wiegand et al., 2019), annotator bias (Ross et al., 2017), gender bias (Binns et al., 2017; Park et al., 2018), racial bias (Sap et al., 2019), political bias (Wich et al., 2020), etc. Dayanik and Padó (2020) focuses on frequency bias of person name in their dataset, while we handle person name considering the passage of time.

In various studies, bias analysis and mitigation are addressed for the fact verification task, where given texts as judged as factual or otherwise from several pieces of evidence, and is one of the recognizing textual entailment tasks. For example, Schuster et al. (2019) and Suntwal et al. (2019) proposed mitigation methods by replacing some words with specific labels to build a robust inference model for out-of-domain data. However, to the best of our knowledge, there has been no study or analysis of bias in fake news detection datasets.

3 Resources

3.1 Datasets

We examine diachronic bias by analyzing four fake news detection datasets with different domains and creation periods. Each article and post in these datasets has a binary label (real/fake). The details of the datasets are as follows and further information is provided in Appendix A:

MultiFC (Augenstein et al., 2019): This is a multi-domain dataset containing over 36,000 headlines from 38 fact-checking organizations. We extracted 7,861 headlines from 2007 to 2015 and regarded headlines labeled as "truth!," "true," or "mostly true" as real; those labeled "mostly false" or "false" as fake.

Horne17 (Horne and Adali, 2017): This dataset contains news articles on the 2016 US presidential election, which are labeled real/fake/satire, based on the investigation of BuzzFeed News. We used articles with fake and real labels.

Celebrity (Pérez-Rosas et al., 2018): This dataset is composed of news articles verified by Gossipcop, which targets news related to celebrities. Most of the articles, whose topics are mainly sensational, such as fights between celebrities, were published between 2016 and 2017.

Constraint (Patwa et al., 2020): This dataset was used in the CONSTRAINT 2021 shared task and consists of social media posts related to COVID-19. These posts were verified by fact-checking sites such as Politifact and Snopes.

3.2 Correlation between phrases and labels

We investigated the correlation between phrases and labels to examine bias in each dataset. To capture high-frequency phrases that are highly correlated with a particular label, we use local mutual information (LMI) (Evert, 2005). Given a dataset D, the LMI between a phrase w and label l is defined as follows: $LMI(w,l) = p(w,l) \cdot \log\left(\frac{p(l|w)}{p(l)}\right)$, where $p(w,l)$ is calculated as $\frac{\text{count}(w,l)}{|P|}$, $p(l|w)$ as $\frac{\text{count}(w,l)}{\text{count}(w)}$, $p(l)$ as $\frac{\text{count}(l)}{|P|}$, and $|P|$ is the number of occurrences of all phrases in D.

Table 1 presents bi-grams that are highly correlated with each label in MultiFC and Constraint[1]. In MultiFC, the headlines prior to 2015 containing words referring to the U.S. president at the time,

[1] Appendix B lists bi-grams with high LMI in Horne17 and Celebrity.

such as "barack obama," were highly correlated with fake labels. In Constraint, words such as "bill gates" and "donald trump," were highly correlated with the fake labels. These results revealed a bias in the relationship between specific person names and labels. The detection model trained on one of these datasets is not adaptable to instances such as the change of president, and thus does not work well on other datasets.

4 Diachronic bias mitigation

4.1 Masking methods

We examine multiple masking methods, starting from word deletion to word replacement for input text data to mitigate the diachronic bias and to build a robust detection model for out-of-domain data. We utilize Named Entity Recognition (NER) (Akbik et al., 2019) in Flair (Akbik et al., 2018) to search for words to be used as masks. Examples of each masking method are listed in Appendix D.

Named Entity (NE) Del: Words tagged with NEs are removed. This masking method aims to build a detection model independent of NE, same as (Suntwal et al., 2019).

Basic NER: Words tagged with NEs are replaced with the corresponding labels, such as PER (person label), LOC (location label), etc.

WikiD: Words tagged with PER labels are replaced with Wikidata (Wikidata) label, specifically, position held (P39), or alternatively occupation (P106) depending on availability. For example, the use of Wikidata at that time made it possible to replace the phrase "barack obama" in articles from 2015 and "donald trump" in those from 2020 with the same label as President of the United States (Q11696). This makes fake news detection models more robust against the passage of time and potentially more effective in mitigating the bias.

WikiD+Del: Words tagged with PER labels are replaced by the same rule as WikiD, and words tagged with other NEs are removed.

WikiD+NER: Words tagged with PER labels are replaced by the same rule as WikiD, and words tagged with other NEs are replaced with the corresponding label.

4.2 Experimental Setting

We verify the effectiveness of masking methods for fake news detection. We examine how well the detection models perform with each of the masking methods against in-domain and out-of-domain for all datasets. Our in-domain experiments mainly investigate the effect of each masking method on accuracy in the same domain. Our out-of-domain experiments validate the effect of each masking method in datasets with considering the flow of time. We consider that out-of-domain setting is close to reality and useful whether each masking method is effective against diachronic bias.

Model: Our experiments utilize a pretrained model, $\mathrm{BERT}_{\mathrm{BASE}}$ model (Devlin et al., 2019), which is made freely available by Google. Labels (LOC, Q11696, etc.) replaced by each masking method were handled as new tokens during the fine-tuning of the pretrained model.

Data and Evaluation: Each dataset is randomly divided into training (80%) and test (20%) sets. The time-based splitting is suitable for our experimental settings than the random splitting in each dataset. However, it was difficult for us to apply the time-based splitting for in-domain experiment because the published time is not described in most of the samples[2]. Out-of-domain experimental settings mean the same verification as the time-based splitting. For example, the evaluation of models, trained in MultiFC consisting of events in 2015, on Constraint consisting of events in 2020 is equivalent to time-based splitting.

4.3 Experimental Results and Discussions

4.3.1 In-domain data

Table 2 presents the accuracies of each masking method against the in-domain data. No Mask, with no application of the masking method, achieved the highest accuracy in all datasets except Constraint. On the other hand, WikiD achieved the highest accuracy on Constraint. In addition, there was only a slight difference in accuracy between No Mask and the other masking methods, even in the datasets where No Mask had the highest accuracy. These results suggest that these masking methods result in insignificant decrease in accuracy when tested with the in-domain dataset.

4.3.2 Out-of-domain data

Table 3 presents the accuracies of each masking method against the out-of-domain data. For almost all of the out-of-domain test data, each masking method achieved higher accuracy than No Mask.

[2] In appendix G, we also conduct an experiment using MultiFC, which contains the published time, with a time-based splitting in train and test sets for trying to remove the effect of domain shift.

Method	Test set			
	MultiFC	Horne17	Celebrity	Constraint
No Mask	**0.681**	**0.746**	**0.760**	0.960
NE Del	0.656	0.706	0.750	0.959
Basic NER	0.659	0.735	0.750	0.950
WikiD	0.675	0.725	0.730	**0.967**
WikiD+Del	0.660	0.706	0.700	0.959
WikiD+NER	0.660	0.640	0.730	0.957

Table 2: Accuracy of each masking method against in-domain data. Bold indicate the highest accuracies.

Train set	Method	Test set			
		MultiFC	Horne17	Celebrity	Constraint
MultiFC	No Mask	-	0.706	**0.660**	0.530
	NE Del	-	0.706	0.590	*0.664
	Basic NER	-	0.725	0.600	*0.680
	WikiD	-	**0.746**	0.590	***0.689**
	WikiD+Del	-	0.725	**0.660**	*0.669
	WikiD+NER	-	0.632	0.520	*0.667
Horne17	No Mask	0.504	-	0.670	0.481
	NE Del	***0.551**	-	**0.680**	*0.553
	Basic NER	0.523	-	0.670	***0.563**
	WikiD	*0.525	-	0.620	0.487
	WikiD+Del	0.523	-	0.610	0.515
	WikiD+NER	0.500	-	0.630	*0.531
Celebrity	No Mask	0.533	0.451	-	0.583
	NE Del	0.545	0.529	-	***0.763**
	Basic NER	0.521	0.549	-	0.568
	WikiD	***0.555**	*0.549	-	*0.724
	WikiD+Del	0.534	0.529	-	*0.663
	WikiD+NER	0.525	***0.568**	-	0.598
Constraint	No Mask	0.542	0.568	0.580	-
	NE Del	0.531	0.588	0.570	-
	Basic NER	0.543	0.568	0.580	-
	WikiD	***0.556**	0.607	0.570	-
	WikiD+Del	0.544	**0.627**	**0.590**	-
	WikiD+NER	0.549	0.607	0.570	-

Table 3: Accuracy of out-of-domain data. The leftmost column lists the training set, and each column with accuracy corresponds to each test set. We applied the statistical significance test by McNemar's test (Dror et al., 2018) with Bonferroni correction to each method compared to No Mask. * indicates the significant difference over No Mask ($p < 0.05$).

Results of No Mask indicate the difficulty of adapting to out-of-domain data. For example, the model trained on Constraint achieved a high accuracy of 0.967 on Constraint (refer to Table 2); however, the model trained on datasets except Constraint achieved low accuracies, ranging from 0.48 to 0.58 on the same test set. These results imply the difficulty of generalizing the fake news detection model. NE Del achieved the same of higher accuracy than No Mask in 9 out of 12 settings, although it is the simplest masking method. In particular, NE Del trained on Horne17 achieved the highest accuracy for MultiFC and Celebrity among models trained on Horne17. Although Basic NER trained on Horne17 also achieved the highest accuracy for Constraint, the improvement was smaller compared to that of NE Del.

Wikidata-based masking methods WikiD and WikiD+Del achieved higher accuracy compared to No Mask in 9 settings, except when testing with Celebrity test set. In particular, the accuracies in WikiD have statistically significant improvements in 6 settings, compared to No Mask. For example, the model trained on MultiFC achieved a significant improvement in accuracy from 0.530 (No Mask) to 0.689 (WikiD), against Constraint test set. These results reveal that the masking method using Wikidata could mitigate the diachronic bias and build robust models even for out-of-domain data. In appendix H, we show some examples, which WikiD accurately classifies while No Mask makes the wrong classification. However, against Celebrity test set, whose domain is entertainment, the accuracy of No Mask is almost the same as that of these masking methods. We consider that this is due to the non-applicability of Wikidata owing to the difference in domain between Celebrity and other datasets; 53.6% of Wikidata labels of Horne17 were found in MultiFC, while only 33.6% of those were found in Celebrity (refer to Appendix E). Additionally, WikiD+NER has comparatively lower accuracies than WikiD+Del. This indicates that we can build a more robust model by removing entities other than person names. We believe that the model can focus on stylistic features by removing extra entity information.

5 Conclusion

This study proposed a new bias concept, Diachronic Bias, caused by the difference in the creation period of various fake news datasets. We firstly examined the deviation of phrase appearance in respective fake news detection datasets with different creation periods. We then proposed masking methods using Wikidata to mitigate the influence of person names. These masking methods achieved higher accuracy in out-of-domain datasets and showed to be made a model more robust.

In the future, more sophisticated approaches such as utilizing a knowledge graph to mitigate diachronic bias would be considered for fake news detection models. In addition to diachronic bias, political bias and racial bias are likely to exist in fake news detection datasets; clarifying these biases in detail is an important next research direction.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proc. of NAACL*, pages 724–728.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proc. of COLING*, pages 1638–1649.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proc. of EMNLP-IJCNLP*, pages 4677–4691.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Proc. of SocInfo*, pages 405–415.

Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in twitter during the 2016 us presidential election. *Nature Commun*, 10(1):1–14.

CNBC. 2013. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked. https://www.cnbc.com/id/100646197.

Erenay Dayanik and Sebastian Padó. 2020. Masking actor information leads to fairer political claims detection. In *Proc. of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proc. of ACL*, pages 1383–1392.

Stefan Evert. 2005. The statistics of word cooccurrences: word pairs and collocations.

Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. 2020. Analyzing temporal patterns of topic diversity using graph clustering. *The Journal of Supercomputing*, pages 1–14.

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 *Proc. of ICWSM*, volume 11.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proc. of EMNLP*, pages 2799–2804.

Parth Patwa, Shivam Sharma, Srinivas PYKL, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Fighting an infodemic: Covid-19 fake news dataset. *arXiv:2011.03327*.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proc. of COLING*, pages 3391–3401.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv:1701.08118*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proc. of ACL*, pages 1668–1678.

Ana Lucía Schmidt, Fabiana Zollo, Michela Del Vicario, Alessandro Bessi, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Anatomy of news consumption on facebook. In *Proc. Natl. Acad. Sci. USA*, volume 114, pages 3035–3039.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proc. of EMNLP-IJCNLP*, pages 3410–3416.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2020. An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv:2005.05710*.

Gautam Kishore Shahi and Durgesh Nandini. 2020. FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In *Workshop Proc. of ICWSM*.

Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining disinformation and fake news: Concepts, methods, and recent advancements. *arXiv:2001.00623*.

Sandeep Suntwal, Mithun Paul, Rebecca Sharp, and Mihai Surdeanu. 2019. On the importance of delexicalization for fact verification. In *Proc. of EMNLP-IJCNLP*, pages 3404–3409.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proc. of the Fourth Workshop on Online Abuse and Harms*, pages 54–64.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proc. of NAACL*, pages 602–608.

Wikidata. https://www.wikidata.org/wiki/Wikidata. (accessed Dec 20, 2020).

A Overview of datasets

The domains and the number of samples for each label in each dataset are listed in Table 4. We do not use the validation set in our experiments because the number of samples in Horne17 and Celebrity is small.

Table 4: Overview of datasets

Dataset	Domain	Year	Real	Fake
MultiFC	Multi	2007–2015	3803	4058
Horne17	Political	2016	128	123
Celebrity	Entertainment	2016–2017	250	250
Constraint	COVID-19	2020	5600	5100

B LMI in Horne17 and Celebrity

Table 5 shows the top LMI-ranked bi-grams that are highly correlated with each label in Horne17 and Celebrity. The result indicates that fake labels in Horne17 and Celebrity have a high correlation with the celebrity's name, such as "brad pitt" and "kate middleton" as well as "donald trump" and "hillary clinton." On the other hand, real labels have a high correlation with common phrases such as "i had" and "would be." These four datasets have a tendency to be highly correlated between person names and fake labels and between common phrases and real labels.

C Wikidata for masking methods

We utilized Wikidata corresponding to the creation time of the articles and posts in each dataset. Specifically, we utilized Wikidata released on January 4, 2016, for MultiFC, January 15, 2018, for Horne17 and Celebrity, and December 28, 2020, for Constraint.

D Examples of outputs by each masking method

Table 6 lists outputs by the masking methods.

E Relationship between Wikidata labels in each dataset

Table 8 presents the coverage rate of Wikidata label between each dataset. Each percentage value represents the number of Wikidata labels in the dataset in the left column covered by Wikidata labels in other datasets. The MultiFC and Horne17 datasets have a high coverage rate because they contain articles and posts on the same political topics. On the other hand, the coverage of the Celebrity dataset is lower than that of the other datasets owing to the entertainment domain. Table 9 presents the top-3 appearance ranked Wikidata label in each dataset.

Horne17					
Real			Fake		
Bi-gram	**LMI**	$p(l\|w)$	**Bi-gram**	**LMI**	$p(l\|w)$
trump has	112	0.82	**donald trump**	605	0.42
national security	106	0.88	**hillary clinton**	440	0.50
would be	104	0.72	i think	292	0.68
people who	92	0.89	united states	258	0.51
transition team	88	1.0	have been	230	0.41
mr. trump	80	0.94	**bill clinton**	208	0.70
smug style	77	1.0	we are	206	0.56
george w.	76	0.90	**hillary clinton's**	187	0.58
republican party	76	0.91	**president obama**	171	0.55
new york	70	0.77	**ted cruz**	149	0.80
Celebrity					
Real			Fake		
Bi-gram	**LMI**	$p(l\|w)$	**Bi-gram**	**LMI**	$p(l\|w)$
i think	233	0.90	has been	343	0.55
i dont́	164	0.95	do think	214	0.80
they were	102	0.70	an insider	199	0.88
i had	100	0.94	**brad pitt**	163	0.63
so i	100	0.92	insider told	157	0.90
but i	87	0.79	may have	128	0.85
we were	87	0.89	**kate middleton**	124	0.88
what i	75	0.87	they are	122	0.51
i love	70	0.92	**the weeknd**	119	0.62
when i	69	0.92	**kanye west**	113	0.56

Table 5: Top 10 LMI ranked bi-grams in Horne17 and Celebrity datasets for real and fake labels with their $p(l|w)$. **LMI** are written as value multiplied by 10^6. Person names are written in bold.

F Implementation Details

Our experimental code is public in https://github.com/hkefka385/mitigation_diachronic_fake.

Hyperparameters Hyperparameters for training our model are as below: learning rate is 1.0×10^{-5}, batch size is 16 and sentence length is 512. We selected the initial value in huggingface library (one of the python libraries) as hyperparameters, based on the empirical rule that fine-tuning works well in our dataset.

Training Efficiency Since our model has the same architecture as BERT (Devlin et al., 2019) except for new tokens that are added for new labels, it holds approximately 110M parameters. We use a Quadro RTX 8000 GPU to train our model. Training our model takes 3 epochs (about 1 hour) for MultiFC and Constraint datasets, and 8 epochs (about 20 minutes) for Horne17 and Celebrity datasets.

No Mask	18 states including **US UK** and **Australia** request PM $\boxed{\textbf{Modi}}$ to head a task force to stop coronavirus
NE Del	18 states including and request PM to head a task force to stop coronavirus
Basic NER	18 states including **LOC LOC** and **LOC** request PM $\boxed{\textbf{PER}}$ to head a task force to stop coronavirus
WikiD	18 states including **US UK** and **Australia** request PM $\boxed{\textbf{Q22337580}}$ to head a task force to stop coronavirus
WikiD+Del	18 states including and request PM $\boxed{\textbf{Q22337580}}$ to head a task force to stop coronavirus
WikiD+NER	18 states including **LOC LOC** and **LOC** request PM $\boxed{\textbf{Q22337580}}$ to head a task force to stop coronavirus

Table 6: Example fake news with the application of masking methods. $\boxed{\text{PER label}}$ is represented by squares and other **NE label** is represented in bold. In Wikidata, Q22337580 indicates Chief Minister of Gujarat.

Dataset	Text	Label
Horne17	**Trump** wins Electoral College vote as insurgency fizzles WASHINGTON - **Donald Trump** will - officially - become president next month. Trump surpassed the 270 electoral votes ...	Real
Constraint	How many FROM covid 19? How many died because New York and New Jersey screwed the elderly?? Thats all **trumps** fault right? When **trump** shut down travelhe a racist **Trump** puts a team together to figure out the virusits not diverse enough	Fake
Constraint	**AG Barr** Suggests an End to the Coronavirus Lockdown URL	Fake

Table 7: Some examples, which WikiD trained on MultiFC accurately classifies while No Mask trained on MultiFC makes the wrong classification. These examples include politicians such as “Donald Trump” and “William Barr”.

	MultiFC	Horne17	Celebrity	Constraint
MultiFC	-	37.4%	29.4%	30.1%
Horne17	53.6%	-	36.4%	39.9%
Celebrity	38.9%	33.6%	-	33.6%
Constraint	35.3%	34.1%	28.6%	-

Table 8: Coverage rate of Wikidata label between each dataset. These percentage values indicate how much of Wikidata labels in the data listed in the left-most column is contained in other datasets.

Rank	MultiFC	Horne17	Celebrity	Constraint
1	President of the U.S.	President of the U.S.	actor	President of the U.S.
2	U.S. representative	Attorney General of Arkansas	singer	CEO
3	Secretary of State	Secretary of State	television actor	Mayor of London

Table 9: Top-3 ranked appearances in Wikidata label of each dataset

G Experiments on time-based splitting of MultiFC

Not only in-domain and out-of-domain experimental settings, but we also conduct an experiment based on time-based splitting of MultiFC into train and test sets. The time-based experimental setting is similar to the intent of out-of-domain experimental settings (refer to Sec.4.3.2). It intends to verify the effect of each masking method in the same dataset for trying to remove more of the effects of domain shift. The experimental setting can not be applied to other datasets because only MultiFC has the published time of each sample. We set samples published in 2012–2015 as train set (3508 samples) and samples published in 2016–2018 as test set (2389 samples).

Method	Test set: MultiFC (2016-2018)
No Mask	0.639
NE Del	0.659
Basic NER	0.639
WikiD	*0.664
WikiD+Del	***0.671**
WikiD+NER	0.644

Table 10: Accuracy of each masking method against MultiFC test set (2016-2018). Bold indicates the highest accuracy.

Table 10 presents the accuracies of each masking method against MultiFC test set (2016-2018). Same as out-of-domain experiments, almost masking methods achieved higher accuracy than No Mask. This show that the masking method using Wikidata could mitigate the diachronic bias and build robust models even for time-based splitting of a dataset.

H Some examples of dataset

In Table 7, we show some examples, which WikiD trained on MultiFC accurately classifies while No Mask trained on MultiFC makes the wrong classification.