Temporal Generalizability in Multimodal Misinformation Detection

Nataliya Stepanova
University of Edinburgh
npstepanova@gmail.com

Björn Ross
University of Edinburgh
b.ross@ed.ac.uk

Abstract

Misinformation detection models degrade in performance over time, but the precise causes of this remain under-researched, in particular for multimodal models. We present experiments investigating the impact of temporal shift on performance of multimodal automatic misinformation detection classifiers. Working with the r/Fakeddit dataset, we found that evaluating models on temporally out-of-domain data (i.e. data from time stretches unseen in training) results in a non-linear, 7-8% drop in macro F1 as compared to traditional evaluation strategies (which do not control for the effect of content change over time). Focusing on two factors that make temporal generalizability in misinformation detection difficult, content shift and class distribution shift, we found that content shift has a stronger effect on recall. Within the context of coarse-grained vs. fine-grained misinformation detection with r/Fakeddit, we find that certain misinformation classes seem to be more stable with respect to content shift (e.g. Manipulated and Misleading Content). Our results indicate that future research efforts need to explicitly account for the temporal nature of misinformation to ensure that experiments reflect expected real-world performance.

1 Introduction

Misinformation proliferation in sectors from public health to politics has shaped public attitudes, undermining trust in reputable organizations and science as a whole. The threat of misinformation is so severe that Lin (2019) qualifies "cyber-enabled information warfare" as an existential risk that can undermine the structure of public discourse, with its potential harm to civilization on par with climate change and nuclear warfare. Although misinformation is not a novel phenomenon, its impact on society has been exacerbated by the advent of social media, which has increased the rate and ease of misinformation spread (Murayama, 2021). As such, automated misinformation detection models are vital in mitigating misinformation's destabilizing effects on society.

In the case of multimodal data, such as an image with a text caption, we must also consider the interaction between modalities (e.g. does the caption contradict the image?), which makes multimodal misinformation detection a harder task (Abdali, 2022). Nevertheless, accounting for multimodality is vital for real-world applications, since a large portion of information shared online is multimodal.

While some Machine Learning (ML) methods can be comparable to human annotators in labelling fake news (Pérez-Rosas et al., 2017), building a multimodal model that is generalizable, explainable and scalable remains a challenge. In the current work, we explore model performance on future, unseen data (temporal generalizability). Undoubtedly, the topics most subject to misinformation change rapidly with time, as does the misinformation itself. As such, a real-world model trained for optimal performance at a specific time point will likely continue to degrade in performance over time. However, most surveyed literature did not directly account for this expected drop in performance, testing models on data collected from the same time period as training data. This practice inflates expected model performance for future applications, referred to by Murayama (2021) as an issue with the model's "velocity".

Thus, we undertook experiments to quantify temporal generalizability of multimodal misinformation detection models. Our goals are twofold:

GOAL 1: Quantify the expected drop in performance when a multimodal classifier trained on the r/Fakeddit dataset is tested on out-of-temporal-domain content.

GOAL 2: Isolate the effect of content shift on the expected performance drop, specifically disentangling trends with reference to the r/Fakeddit misinformation classes.

Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP, pages 76–88
December 6, 2023 ©2023 Association for Computational Linguistics

2 Literature Review

2.1 Detecting Misinformation

Works such as Murayama (2021) and Wardle et al. (2018) provide an overview of different definitions of “fake news” and related concepts (misinformation, rumor, satire, propaganda, etc.). But the question of how to define “misinformation” is still an on-going area of research. For example, Abdali (2022) listed binary ground truth (true vs. false) and lack of granularity in labels in existing datasets as a data-related challenge in the field. To avoid the issue of coarse labeling, some researchers formulate the misinformation detection task as a regression problem. For our purposes, we use “misinformation” interchangeably with “fake news” to refer to content that contains some element of untruth, regardless of intention.

Fake news classification research is closely related to tasks such as fact verification, fact checking, rumor/stance detection, and sentiment extraction (Oshikawa et al., 2020). A classic ML fake news detection model represents input content (e.g. text and/or image) with manually selected features and feeds these into a classification or regression model (Zhou and Zafarani, 2020). The advent of powerful deep learning models like BERT (Devlin et al., 2019) for text or ResNet (He et al., 2016) for images made it possible to step away from manually crafted features toward learned representations extracted from hidden layers.

For multimodal information, the main question lies in whether to process each individual modality and then combine the predictions (ensemble methods), or whether to extract cross-modal features that account for inter-modal interactions (Abdali, 2022). In ensemble methods, one can fuse the modalities at the raw input level – “early fusion”, after extracting modality-specific features (e.g. through vector concatenation) – “intermediate fusion”, or instead fuse predictions of each modality-specific model – “late fusion” (Boulahia et al., 2021). Examples of cross-modal features used include measures of similarity between the input text and images (Giachanou et al., 2020). Another explored avenue for extracting cross-modal interactions consists of using attention mechanisms. However, attention-based models are not as easily explainable, with regions to attend to often discovered through trial and error (Abdali, 2022).

2.2 Evaluation and Temporal Generalizability

The problem of model generalizability to unseen data is a known challenge in ML, often addressed through methods such as domain adaptation and transfer learning (e.g. see Kouw and Loog (2018)). Works such as Suprem et al. (2019) and Žliobaitė (2010) have used continuous/incremental learning to train models to respond to “concept drift”. In the context of misinformation detection, what is defined as "in-domain" and "out-of-domain" can vary. Experiments posed by Nan et al. (2021) and Min et al. (2022) define "domain" as "subject/topic of data", e.g., training models on political sources and testing on social content. In this paper, we treat different time periods as different "domains", since content even within the same subject/topic changes so rapidly. We refer to a model’s ability to perform on data from a time period not seen in training as its “temporal generalizability".

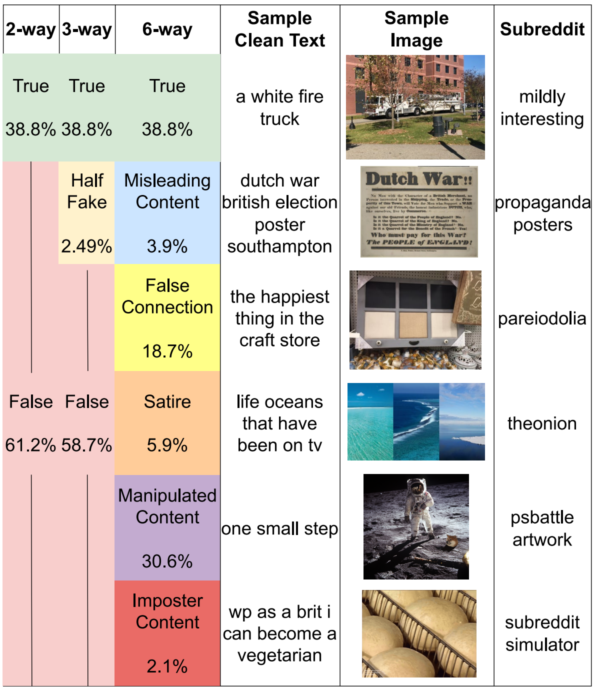
Bozarth and Budak (2020) explored such “temporal generalizability" of misinformation detection models by comparing evaluation strategies: **classic** (common N-fold cross-validation across the entire dataset), **forecast** (evaluating on future data from a time period past the end of training), **bydomain** (evaluation against content not seen in training). They found that “classic” evaluation (which was most often encountered in surveyed literature) yielded higher performance than “forecast” evaluation (which closely mimics what happens with production models). Horne et al. (2019) similarly found that performance of fake news classifiers worsens over time, although they note that certain features (e.g. content-based features like style of writing) are more robust to temporal changes. Alkhalifa et al. (2023) observed a more pronounced model deterioration with time for “open-domain” content (e.g. social media) as opposed to “closed-domains” (e.g. book reviews).

Improving temporal generalizability has been explored on text-only models: Zhu et al. (2022) used an “entity debiasing framework”, Suprem and Pu (2022) proposed a new method based on K-Means clustering, while Murayama et al. (2021) showed that using masking during text-based model training resulted in a better generalization accuracy. The generalizability of multimodal models has not been explored to the same extent as for unimodal (specifically text-only) models. Moreover, to the best of our knowledge, no prior temporal generalizability study has used the r/Fakeddit dataset.

3 Methods

3.1 Data: r/Fakeddit

We used the r/Fakeddit dataset, introduced by Nakamura et al. (2020) as a “multimodal benchmark dataset for fine-grained fake news detection”. Compared to other multimodal datasets, it is one of the largest publicly available,[1] and contains both binary and fine-grained labels. Data was sampled from 22 subreddits (refer to Table 8 in the Appendix for specifics). Of its 1 million samples, roughly 650K are multimodal, containing text (title of Reddit post) and an associated image. These span June 1, 2008 to November 15, 2019 and are the focus of our investigations. Labels consist of three levels of granularity: 2-, 3- or 6-way (see Figure 1).



2-way	3-way	6-way	Sample Clean Text	Sample Image	Subreddit
True 38.8%	True 38.8%	True 38.8%	a white fire truck		mildly interesting
False 61.2%	Half Fake 2.49%	Misleading Content 3.9%	dutch war british election poster southampton		propaganda posters
	False 58.7%	False Connection 18.7%	the happiest thing in the craft store		pareiodolia
		Satire 5.9%	life oceans that have been on tv		theonion
		Manipulated Content 30.6%	one small step		psbattle artwork
		Imposter Content 2.1%	wp as a brit i can become a vegetarian		subreddit simulator

Figure 1: Example r/Fakeddit data for all possible 2-, 3-, and 6-way classes with relative class sizes.

r/Fakeddit is imbalanced, with the imbalance getting more pronounced as the classification gets more fine-grained. In the 2-way labels, 255,913 (38.81%) of posts are labelled as True and 403,451 (61.19%) False. The 3-way labeling is roughly the same as for 2-way labeling, just a portion of the Fake samples is listed as Half Fake (2.49% of data). The most fine-grained 6-way labeling, with 5 sub-classes for fake content, contains the most imbalances. In this report, we focus only on 2-way and 6-way classification, as our 3-way models behaved extremely similarly to the 2-way models, likely due to the relatively small size of the Half Fake class.

Preprocessing We keep the pre-processed text of Nakamura et al. (2020). Images were preprocessed following the practices of He et al. (2016), Krizhevsky et al. (2012), and Simonyan and Zisserman (2015): we resized and randomly cropped images to force image dimensions to 224x224, then normalized each pixel value using the mean and standard deviation of RGB values in the ImageNet dataset (default in Pytorch).[2]

3.2 Train-Validation-Test Data Splits

We prepared three train-val-test splits, changing the temporal range of information available to models.

Original (OG) Data Split: A random partitioning of the dataset into train, val, and test sets provided by the r/Fakeddit authors. All three spanned the entire decade of available data (see Table 1).

Split	Time Covered	# Posts	% Data
Train	06.2008-11.2019	544,288	82.56%
Val	07.2008-10.2019	57,551	8.72%
Test	06.2008-10.2019	57,525	8.72%

Table 1: Original train-val-test split statistics.

Temporal Data Split: All three splits cover a separate time period. We sorted all available data by creation timestamp and separated it into three consecutive chunks corresponding in size to the OG train, val, and test splits (to control for dataset size). Splitting the data this way ensured that models would be both validated and evaluated on temporally out-of-domain data.

Split	Time Covered	% Data	Months
Train	06.2008-04.2019	82.56%	131
Val	04.2019-07.2019	8.72%	3
Test	07.2019-11.2019	8.72%	4

Table 2: Temporal train-val-test split statistics.

Multiple Test Splits Over Time: Has 5 consecutive test splits, designed to quantify the change in performance as the test set gets further removed from the training data. Although the raw count of data points in our train and validation sets had to

[1] Refer to https://github.com/entitize/Fakeddit

[2] https://pytorch.org/vision/stable/transforms.html#scriptable-transforms

decrease, we controlled for the *relative* proportion (82% to 9%) of samples in them (see Table 3). We made the test splits roughly equal to that of the validation set. We controlled for the temporal coverage (not size) of the test sets because in a hypothetical real-life scenario it would be preferable to know how soon after deployment (not after how many runs) a model should be retrained.

Split	Time Covered	% of Data	Days
Train	06.2008-05.2017	54.05%	3,287
Val	06.2017-11.2017	5.71%	155
Test 1	11.2017-04.2018	5.97%	153
Test 2	05.2018-09.2018	7.57%	155
Test 3	09.2018-02.2019	2.48%	155
Test 4	02.2019-07.2019	14.02%	152
Test 5	07.2019-11.2019	10.19%	128

Table 3: Data split stats for multiple test splits.

3.3 Model Architecture and Training

Our multimodal models followed one version of the "ensemble method" described by Abdali (2022):

1. Text and image were processed individually to extract modality-specific features
2. Feature vectors were concatenated ("intermediate fusion", see Boulahia et al. (2021))
3. The resulting concatenated vector was fed into a neural network classifier

To process cleaned submission titles, we used a variant of the popular transformer-based pre-trained language model BERT (Devlin et al., 2019). BERT has successfully been used to embed input for misinformation detection models (see Nakamura et al. (2020) and Segura-Bedmar and Alonso-Bartolome (2022)). We used a pre-trained RoBERTa model (variant *all-distilroberta-v1*)[3] to obtain 768-dim embeddings for sample text.

We used a pre-trained network called ResNet-50 to extract image features (He et al., 2016). The ResNet architecture won the ILSVRC 2015 image classification task and has since remained a popular backbone for computer vision models. It has also made its way into misinformation classifiers, demonstrating potential for transfer learning. For example, Nakamura et al. (2020) found that using ResNet-50 for classification with r/Fakeddit resulted in a better performance than using VGG16 (Simonyan and Zisserman, 2015) or EfficientNet (Tan and Le, 2019), both alternative deep convolutional neural networks (CNNs) commonly used for computer vision. We similarly used ResNet-50's penultimate layer weights to represent each input image with 2048-dim vectors.

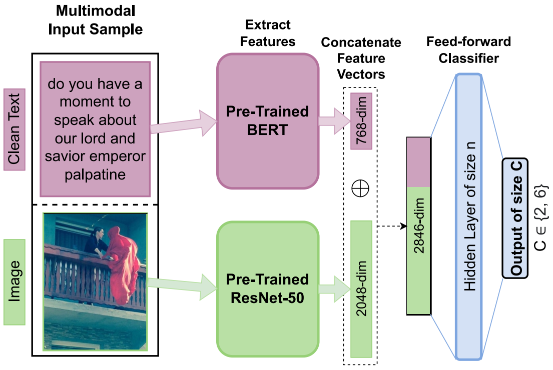


Figure 2: Ensemble multimodal model for 2-way and 6-way classification. The best-performing hidden layer size, *n*, was found through hyperparameter tuning.

Text and image features were combined through simple concatenation. This concatenated vector was fed through a one-hidden-layer feed-forward neural network for classification (see Figure 2).

To train our models, we first loosely followed both Nakamura et al. (2020) and Segura-Bedmar and Alonso-Bartolome (2022) to choose hyperparameters. We used cross-entropy loss and the Adam optimizer. Each model was selected after conducting extensive hyperparameter tuning over hidden layer size (n) and learning rate (lr). We tested all possible pairs of the following: $n = 2^i, i \in \{5, 6, 7, 8, 9, 10, 11, 12, 14\}$, and $lr \in \{0.01, 0.001, 0.0001, 0.00001\}$. We used batches of size 256 and trained for a max of 20 epochs, with early stopping where validation accuracy did not improve over 4 consecutive epochs. Each final model was selected by choosing the hyperparameter setting that maximized accuracy on the validation set (see Appendix for specifics).

3.4 Evaluation

In an imbalanced class setting, a micro F1 score can be inflated by high performance on high frequency classes, whereas macro F1 is a better reflection of model performance across all classes, regardless of size. We use both metrics to evaluate our models for Experiment 1. For Experiment 2, we report micro F1 change over time, as that is representative of real-world performance after deployment.

[3] https://huggingface.co/sentence-transformers/all-distilroberta-v1

4 Experiment 1: OG vs. Temporal

Our first experiment quantified the drop in performance when evaluating on temporally "out-of-domain" data (GOAL 1). We built multimodal models using the original vs. temporal splits for 2-way and 6-way classification. For each type of prediction and data split, we evaluated a Baseline model that randomly classified test samples proportionally to their rate of occurrence in the training set.

4.1 Results: 2-Way Classification

We report micro and macro F1 (see Table 4) and confusion matrices (see Figure 3) both for the model trained on the original train-val-test split and the temporal split. Our confusion matrices are normalized over the True/Actual labels (all rows sum to 1.0), so entries along the main diagonal represent recall per class.

	Trained		**Baseline**	
	OG	**Temp.**	**OG**	**Temp.**
Micro F1	0.85	0.81	0.52	0.56
Macro F1	0.85	0.78	0.50	0.50

Table 4: Exp. 1 evaluation metrics for 2-way models.

When the train-val-test split was changed to be temporal, we saw a 4% decrease in Micro F1 and 7% decrease in Macro F1. Both OG and Temporal models outperformed their respective Baselines.

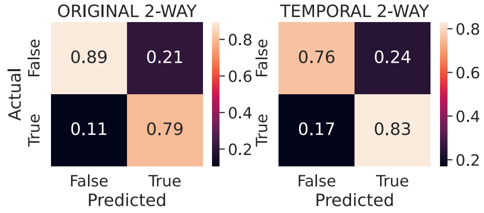


Figure 3: Confusion matrices: OG vs. temporal 2-way.

Looking at the confusion matrices, per-class recall dropped 13% for the Fake class and increased 4% for the True class. The Temporal model was generally predicting more samples into the True class (values in column 2 for both rows are greater than column 1). Detection of Fake samples worsened more than that of True samples.

4.2 Results: 6-Way Classification

We report micro and macro F1 scores in Table 5 and confusion matrices (normalized over the True labels) for 6-way models, both for the original (see Figure 4) and temporal (see Figure 5) splits.

	Trained		**Baseline**	
	OG	**Temp.**	**OG**	**Temp.**
Micro F1	0.76	0.72	0.29	0.28
Macro F1	0.60	0.52	0.17	0.17

Table 5: Exp. 1 evaluation metrics for 6-way models.

When the train-val-test split was changed to temporal, the 6-way model drop in performance was similar to the 2-way models. Micro F1 dropped by 4% and macro F1 by 8%. Both OG and Temporal models, nevertheless, performed substantially better than their respective Baselines. The Temporal model performed worse on lower frequency classes, hence macro F1 was affected more than micro F1.

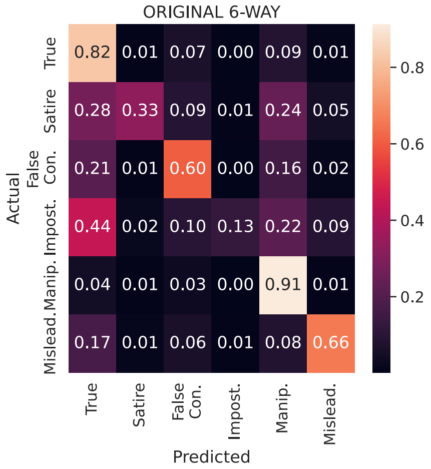


Figure 4: Confusion matrix for original 6-way model.

The OG 6-way model achieved the best per-class recall on True and Manipulated Content classes, potentially since they comprise the majority of training data (39% and 31%, respectively). Perhaps more surprising was the model's ability to achieve 66% recall on Misleading Content, which comprises only 4% of the training set (21K samples). The worst per-class recall performance was on Imposter Content (13%). 44% of True samples was predicted to Imposter Content, making it the most evasive misinformation class (followed by Satire at 28% misclassification to True). Perhaps this type of data is hard to detect, or there were simply not enough samples for the model to learn (11K samples for Imposter Content and 32K for Satire).

The temporal split decreased per-class recall on almost all classes but Satire (7% increase) and Imposter Content (2% increase). We also observed a general trend of predicting most samples into the True class, regardless of the actual label (see the

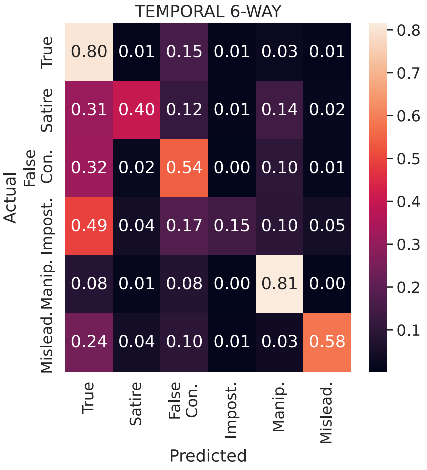


Figure 5: Confusion matrix for temporal 6-way model.

increase across the entire first column) – likely because the True class comprised 69% of the testing set as opposed to 33% of training. The best recall performance was again achieved by the True and Manipulated Content classes. This suggests that unlike the other types of Fake samples, Manipulated images are easiest to detect over time. This makes intuitive sense, as while the subject of photoshopped images might change over time, photoshopping techniques remain relatively stable.

5 Experiment 2: Multiple Test Splits

Our second experiment delved even further into GOAL 1. We quantified the rate of decay in model performance by increasing the number of test splits to five. We probed at the reasons for change in performance (GOAL 2) by comparing our resulting model (**Exp. 2 Normal**) against two variations. There are two reasons performance could change:

1. **Content shift**; e.g. the subject of posts from Apr. 2018 is different than Feb. 2019

2. **Class distribution shift**; e.g. the distribution of True vs. Fake posts changes over time

To isolate the effects of content shift, we evaluated on subsampled sets of the 5 test splits, enforcing the same class distribution and controlling for its effect (**Exp. 2 Balanced**). To isolate the effects of class distribution, we created a **Dummy** classifier that predicted proportionally to class distributions observed in training. Since the dummy classifier did not use content features for prediction, any observed effects over the 5 test splits reflected the class distribution's effect on performance.

5.1 Results: 2-way Classification

We present micro F1 for all three compared models across the 5 test splits for 2-way classification in Table 6. We follow by a per-class, per-test-split, per-model breakdown of F1, precision, and accuracy scores in Figure 6, isolating the effects of content vs. class distribution shift on each metric.

	Model + Evaluation		
	Exp. 2 Normal	**Exp. 2 Balanced**	**Dummy**
Test 1	0.82	0.60	0.76
Test 2	0.79	0.59	0.70
Test 3	0.48	0.46	0.48
Test 4	0.48	0.45	0.50
Test 5	0.40	0.42	0.47

Table 6: 2-way accuracy (micro F1). Exp. 2 Normal represents "real-life" performance, Exp. 2 Balanced isolates content shift effects, and Dummy isolates class distribution shift effects.

In Exp. 2 Normal, accuracy decreased with time, dropping dramatically after Test Split 2. Starting from Test Split 3, it performed worse than a Dummy model. The falling performance of both Exp. 2 Balanced and Dummy models suggest that the drop is due to both content shift and a change in class distributions. Since Test Split 3 starts ∼450 days from the end of training (∼300 days from val), in a real-life scenario our models would likely need to be retrained about once a year.

Looking at Figure 6, we can observe the effects of content shift in column 2 (Exp. 2 Balanced). Recall was unaffected for both True and False classes (compare bottom tiles in columns 1 vs. 2), while precision and F1 fell for both. We turn to column 3 for the effects of class distribution shift (Dummy performance). The relative percentage of the False class in the test sets decreased substantially across the test splits: 76% for Test Split 1, to 73%, 41%, 37%, and finally 31% for Test Split 5. Recall was unaffected by class distribution, whereas precision decreased as the Fake class got smaller.

Overall, the trends in precision of Exp. 2 Normal were most similar to that of the Dummy model, suggesting that class distribution shift has a stronger impact on precision than recall. The trends in recall of Exp. 2 Normal were most similar to that of Exp. 2 Balanced, suggesting that content distribution shift has a stronger effect on recall rather than precision. These two effects combine to give a cumulative negative effect on F1 – worsening

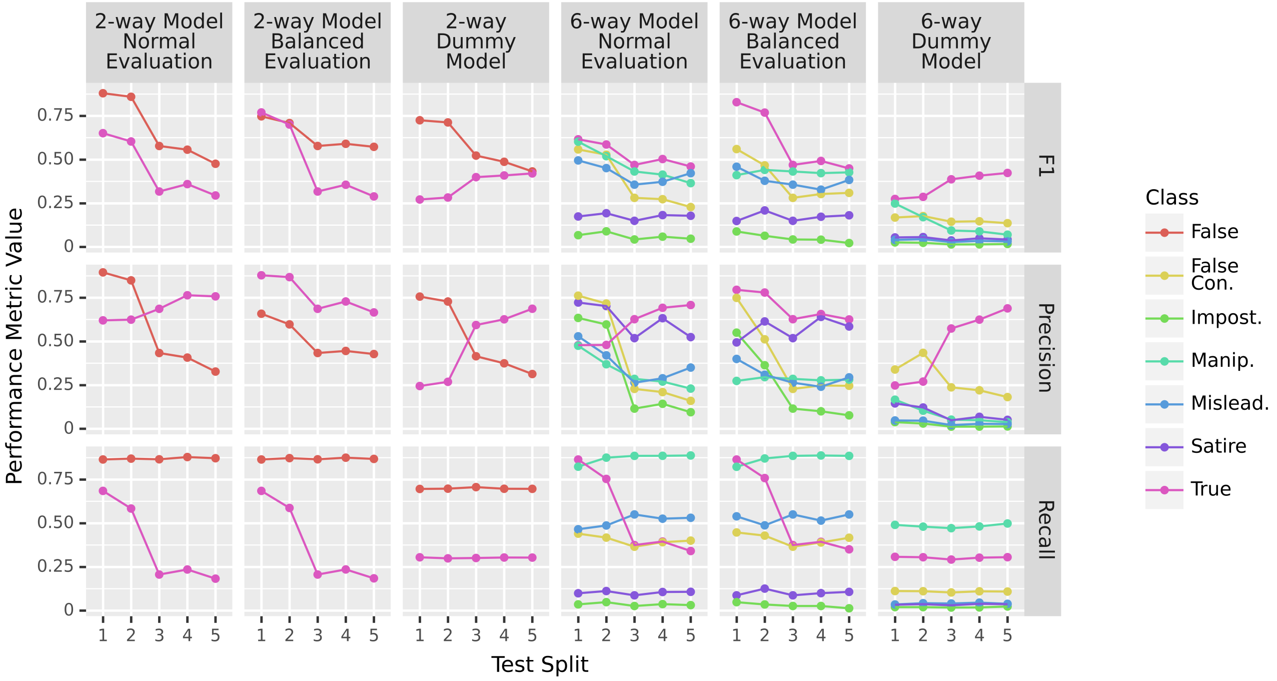


Figure 6: Precision, recall and F1 for 2-way and 6-way models over 5 temporal tests splits. "Normal" model evaluation reports the results as they would be observed in a real-life scenario, whereas "Balanced" model evaluation isolates the effects of content shift and Dummy isolates the effects of class distribution shift.

precision due to class distribution shift and worsening recall due to content shift. Interestingly, the observed per-class and per-metric trends in Exp. 2 Normal can be roughly seen as a sum of the trends in Exp. 2 Balanced and the Dummy model.

5.2 Results: 6-way Classification

We present 6-way micro F1 for all three compared models in Table 7. A breakdown of F1, precision and accuracy scores is again in Figure 6.

	Model + Evaluation		
	Exp. 2 Normal	Exp. 2 Balanced	Dummy
Test 1	0.54	0.20	0.68
Test 2	0.51	0.19	0.62
Test 3	0.38	0.24	0.38
Test 4	0.40	0.24	0.40
Test 5	0.36	0.25	0.37

Table 7: 6-way accuracy (same as micro F1).

The 6-way pattern of accuracy change in Exp. 2 Normal and Dummy was very similar as for 2-way models. Performance fell the most between Test Splits 2 and 3 (to an approximately at-chance performance – see Exp. 1 Baselines). Unlike 2-way classification, the Dummy model outperformed the trained Exp. 2 Normal model from Test Split 1, suggesting that finer-grained misinformation classification may be more temporally unstable than coarser-grained. After the Test Split 3, performance remained relatively stable. However, the Exp. 2 Balanced model performed abysmally through all test splits, starting from the first one, and there was not much change throughout the test splits. This suggests that the drop in performance in the Normal model can be mostly attributed to class distribution shift and not content shift.

Looking at Figure 6, the Exp. 2 Normal model was more similar to Exp. 2 Balanced than the Dummy model with respect to Recall, again suggesting that content shift affects recall. The change in precision for the Dummy model went hand-in-hand with how the relative class distributions changed (the True class got relatively larger with each successive test split, the False Connection class peaked at Test Split 1 and then fell along with the rest, just like the precision values changed).

With Exp. 2 Balanced, True and False Connection classes fell systematically across all splits. However, Manipulated and Misleading content classes performed relatively stably. This either suggests that, potentially, the features our models learned to identify these misinformation classes persist more stably over time than others.

6 Discussion

6.1 Our Contributions

In the r/Fakeddit dataset, a temporal data split resulted in a **4% drop in macro F1 and 7-8% drop in micro F1 for 2-way and 6-way multimodal models** (compare GOAL 1).

For GOAL 2, to isolate the effect of content shift on the performance drop, we found that **content shift seems to affect recall more than precision**. Additionally, **finer-grained misinformation classes do not behave in the same way with regards to temporal generalizability**. Notably, Manipulated and Misleading content classes seemed to be more stable.

Our results for Experiment 1 underline the importance of considering the performance drop in misinformation classification models on a new temporal domain. Our results for Experiment 2 further isolate a period of time where performance drops substantially, suggesting that models may suffer from a sudden and dramatic decrease in performance (as opposed to a gradual worsening of classification accuracy). Specifically in reference to the r/Fakeddit dataset, it seems like there was a qualitative change in content posted between September 2018 and February 2019, where we see the sudden drop in performance (to Baseline levels). Investigating whether this change is due to a specific singular event or has to do with general content shift over time is out of the scope of this paper. In the real-world, guidance from social scientists and/or political scientists who are aware of current online discourse would help identify periods of time when content is expected to change, affecting the performance of deployed models.

Our findings in Experiment 2 with respect to disentangling the effects of content vs. class distribution shift underline the importance of accounting not only for how content might change over time, but also how models will perform in varying class distribution settings. Throughout the experimentation process, we found that when we split the r/Fakeddit dataset temporally, the variation in relative class distribution varied widely across the temporal splits. It is unclear whether this has to do with the way the authors of r/Fakeddit collected the data, or with the underlying distribution of content on Reddit in general. Regardless, in the real world it is very possible that the data collected at time point *X* will vary widely from the data collected at time point *Y*. As such, researchers have to explicitly prepare for how their models will perform in different class distributions settings.

It makes theoretical sense why recall is not affected by class distribution shift and therefore is a useful metric for isolating the effects of content shift. Recall is equal to $\frac{TP}{TP+FN}$, where TP = true positives and FN = false negatives for a certain class. Assuming the content distribution stays the same, an α increase in total data points of a class will correspond to an analogous α increase in both TP and FN, and the scaling factor will cancel out in the recall calculation. The same does not apply to Precision, which is $\frac{TP}{TP+FP}$, with FP = false positives. Whereas TP and FN came from the same class, FP are by definition from a different one, therefore any scaling effects of the two different classes will compound rather than cancel.

6.2 Implications for Research and Industry

As our experiments showed, deployed models can expect a sudden and significant drop in performance, indicating that future research efforts need to explicitly account for the temporal nature of misinformation to ensure that experiments reflect expected real-world performance. Although there is existing research in this space (e.g. Chen and Hasan (2021) look at the temporal generalizability of COVID-19 misinformation detection models), many studies do not account for content change over time. A deeper analysis of why model performance was not always generalizable was hindered by a lack of understanding of what our models were learning. To that end, further research should also look at how models learn what is fake, and whether it is possible to make the decision-making process less dependent on temporal context.

Approaches to misinformation detection can be separated into categories based on how they learn. Zhou et al. (2020) and Shiao and Papalexakis (2021) discuss four: content-based, propagation-based, knowledge-based and source-based models. Our models were implicitly operating off of a content-based approach, relying on latent text and image features to embed samples into a semantic space representing truth-value. Perhaps knowledge-based models are more generalizable, but only if the knowledge base is iteratively updated with the passage of time. Further research would benefit from considering what types of updating (e.g. feature extraction or knowledge base) would be most feasible from an industry perspective.

7 Limitations

Our multimodal models perform worse than related works on r/Fakeddit (Nakamura et al., 2020; Segura-Bedmar and Alonso-Bartolome, 2022). We tried exactly recreating the architecture of Nakamura et al. (2020), but model performance was still lower. Noting that our investigation would benefit from being comparable to related studies, we attempted to locate the source of the discrepancy in performance but were unsuccessful. We think that our difficulty with reproducing existing results is not uncommon, and future research would benefit greatly from studies that explicitly outline guidelines on how to exactly reproduce their architecture (as in Zaeem et al. (2020), Shu et al. (2019)).

Due to computational limitations, we did not fine-tune the feature extraction process on our dataset, instead relying on pretrained RoBERTa and ResNet-50. Building models that are specifically tailored to misinformation datasets for feature extraction might increase performance, though it is unclear if this would change the impact of a temporal data split. Future research could explore whether temporal generalizability is largely dependent on the dataset being used, and whether the obtained results would be different if another dataset was analyzed instead.

Additionally, we did not extensively investigate what validation strategy would work the best for temporal generalizability. Instead, we naively separated the training and validation set temporally from each other, and used that for hyperparameter tuning. Further work could look into training methods specifically designed for maximal temporal generalizability in misinformation detection.

We study the temporal generalizability of multimodal misinformation detection models in one specific language, for one specific platform. Although the experiments presented in this paper are inspired by real-world applications (e.g. deploying a misinformation detection model on a social media platform), it is worth noting that the r/Fakeddit dataset contains some particularities that make it difficult to generalize to broader misinformation detection in other languages and settings. Some of the samples labeled as "Fake" are harmless (e.g. certain memes), although technically they are "untrue" or "manipulated". This raises the question of whether the results presented here are generalizable to datasets that focus on more "serious" topics of information (e.g. COVID-19 or certain political topics). For example, some photoshopped images are evidently meant to entertain, and, although technically they constitute "misinformation", it seems unintuitive to seriously treat them as such. This raises the research question of how to effectively define "misinformation" that both makes sense semantically and also is maximally useful for automated models deployed in the real world, dealing with topics of substantial weight.

8 Ethical considerations

We use the existing r/Fakeddit dataset. Since this was released as a benchmarking dataset for misinformation detection models, our use of this dataset is consistent with the use cases it was intended for. The dataset may contain personal data or offensive content so we ensure that the examples reported in this paper do not make any individuals identifiable or include offensive content. We were not able to find information on the licence associated with this dataset but since it was released for the purpose of benchmarking we assume that our use of this dataset is acceptable.

We study when misinformation detection systems fail to perform well. A malicious actor could potentially exploit this knowledge to decide what kind of misinformation to spread, however, we believe that our results will be far more useful to those who are hoping to improve the temporal generalizability of their systems.

References

Sara Abdali. 2022. Multi-modal misinformation detection: Approaches, challenges and opportunities. *arXiv preprint arXiv:2203.13883*.

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60(2):103200.

Said Yacine Boulahia, Abdenour Amamra, Mohamed Ridha Madi, and Said Daikh. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):1–18.

Lia Bozarth and Ceren Budak. 2020. Toward a better performance evaluation framework for fake news classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):60–71.

Yuanzhi Chen and Mohammad Hasan. 2021. Navigating the kaleidoscope of COVID-19 misinformation using deep learning. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing, pages 6000–6017, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. 2020. Multimodal fake news detection with textual, visual and semantic information. In *International Conference on Text, Speech, and Dialogue*, pages 30–38. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Benjamin D Horne, Jeppe Nørregaard, and Sibel Adali. 2019. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(1):1–23.

Wouter M Kouw and Marco Loog. 2018. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Herbert Lin. 2019. The existential threat from cyber-enabled information warfare. *Bulletin of the Atomic Scientists*, 75(4):187–196.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*, pages 1148–1158.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: A survey. *arXiv preprint arXiv:2111.03299*.

Taichi Murayama, Shoko Wakamiya, and Eiji Aramaki. 2021. Mitigation of diachronic bias in fake news detection dataset. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 182–188, Online. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347.

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.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Isabel Segura-Bedmar and Santiago Alonso-Bartolome. 2022. Multimodal fake news detection. *Information*, 13(6).

William Shiao and Evangelos E Papalexakis. 2021. Ki2te: Knowledge-infused interpretable embeddings for covid-19 misinformation detection. In *KnOD@ WWW*.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Abhijit Suprem, Aibek Musaev, and Calton Pu. 2019. Concept drift adaptive physical event detection for social media streams. In *World Congress on Services*, pages 92–105. Springer.

Abhijit Suprem and Calton Pu. 2022. Evaluating generalizability of fine-tuned models for fake news detection. *arXiv Preprint posted online May*, 15.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Claire Wardle et al. 2018. Information disorder: The essential glossary. *Harvard, MA: Shorenstein Center on Media, Politics, and Public Policy, Harvard Kennedy School*.

Razieh Nokhbeh Zaeem, Chengjing Li, and K Suzanne Barber. 2020. On sentiment of online fake news. In *2020 IEEE/ACM International Conference on*

Advances in Social Networks Analysis and Mining (ASONAM), pages 760–767. IEEE.

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

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery.

Indrė Žliobaitė. 2010. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*.

A Appendix

This appendix first provides additional implementation details, specifically optimal found hyperparameter settings. We follow with details on the subreddits in the dataset and the performance of the 6-way models broken down by subreddit.

As described in section 3.3, we conducted hyperparameter tuning over hidden layer size (n) and learning rate (lr), testing all possible pairs of the following: $n = 2^i, i \in \{5, 6, 7, 8, 9, 10, 11, 12, 14\}$, and $lr \in \{0.01, 0.001, 0.0001, 0.00001\}$. Each final model was selected by choosing the hyperparameter setting that maximized accuracy on the validation set, with optimal learning rate being 0.0001 across the board, $n = 16384$ for OG 2-way and $n = 8192$ for OG 6-way, $n = 8192$ for Temporal 2-way and $n = 1024$ for Temporal 6-way.

The final choice of subreddit and associated truth values went through a rigorous multi-step quality assurance process to justify the use of subreddit-level labels (as opposed to labeling each sample individually), see Nakamura et al. (2020) for a detailed overview of this process and Table 8 for the labels assigned to each subreddit.

Additionally, since all samples from a specific subreddit received the same label (a type of domain-level ground truth, where the domain a sample comes from determines its truth value), refer to Table 9 for a per-subreddit breakdown of 6-way classification accuracy for the original vs. temporal models.

	Label		
subreddit	**2-way**	**3-way**	**6-way**
mildlyinteresting	True	True	True
photoshopbattles	True	True	True
nottheonion	True	True	True
upliftingnews	True	True	True
neutralnews	True	True	True
usanews	True	True	True
pic	True	True	True
usnews	True	True	True
fakealbumcovers	Fake	Fake	Satire
theonion	Fake	Fake	Satire
satire	Fake	Fake	Satire
waterfordwhispersnews	Fake	Fake	Satire
propagandaposters	Fake	Half Fake	Misleading Content
fakefacts	Fake	Fake	Misleading Content
savedyouaclick	Fake	Fake	Misleading Content
psbattle_artwork	Fake	Fake	Manipulated Content
pareidolia	Fake	Fake	False Connection
fakehistoryporn	Fake	Fake	False Connection
misleadingthumbnails	Fake	Fake	False Connection
confusing_perspective	Fake	Fake	False Connection
subredditsimulator	Fake	Fake	Imposter Content
subsimulatorgpt2	Fake	Fake	Imposter Content

Table 8: 2-way, 3-way, and 6-way subreddit-level labels for r/Fakeddit (every sample from a specific subreddit is labeled the same way).

		Per-Subreddit Accuracy	
subreddit	**6-way Label**	**Original**	**Temporal**
mildlyinteresting	True	88.00	80.29
photoshopbattles	True	80.54	83.13
nottheonion	True	89.68	91.26
upliftingnews	True	93.85	94.28
usanews	True	90.29	94.30
pic	True	64.66	71.64
usnews	True	91.16	91.18
neutralnews	True	89.88	NA
fakealbumcovers	Satire	55.56	57.06
theonion	Satire	45.83	35.51
satire	Satire	25.57	19.78
waterfordwhispersnews	Satire	25.00	20.00
pareidolia	False Connection	60.81	60.59
fakehistoryporn	False Connection	74.62	66.18
misleadingthumbnails	False Connection	46.40	54.85
confusing_perspective	False Connection	31.10	33.16
propagandaposters	Misleading Content	75.66	79.72
savedyouaclick	Misleading Content	47.04	46.53
fakefacts	Misleading Content	NA	0.00
subredditsimulator	Imposter Content	29.48	36.51
subsimulatorgpt2	Imposter Content	15.38	7.22
psbattle_artwork	Manipulated Content	87.24	86.52

Table 9: Percent of correctly classified samples per subreddit for original vs. temporal BERT 6-way models.

<table>
<tr><th colspan="12">Motivation</th></tr>
<tr><td colspan="3">Practical
□ △</td><td colspan="3">Cognitive</td><td colspan="3">Intrinsic</td><td colspan="3">Fairness</td></tr>
<tr><th colspan="12">Generalisation type</th></tr>
<tr><td colspan="2">Compositional</td><td colspan="2">Structural</td><td colspan="2">Cross Task</td><td colspan="2">Cross Language</td><td colspan="2">Cross Domain
□ △</td><td colspan="2">Robustness</td></tr>
<tr><th colspan="12">Shift type</th></tr>
<tr><td colspan="3">Covariate</td><td colspan="3">Label</td><td colspan="3">Full</td><td colspan="3">Assumed
□ △</td></tr>
<tr><th colspan="12">Shift source</th></tr>
<tr><td colspan="3">Naturally occuring
□ △</td><td colspan="3">Partitioned natural</td><td colspan="3">Generated shift</td><td colspan="3">Fully generated</td></tr>
<tr><th colspan="12">Shift locus</th></tr>
<tr><td colspan="3">Train–test
□ △</td><td colspan="3">Finetune train–test</td><td colspan="3">Pretrain–train</td><td colspan="3">Pretrain–test</td></tr>
</table>

Table 10: GenBench evaluation card for Exp. 1 (□) and Exp. 2 (△).