
Proceedings Paper:
Mu, Y., Song, X., Bontcheva, K. et al. (2024) Examining the limitations of computational rumor detection models trained on static datasets. In: Calzolari, N., Kan, M-Y., Hoste, V., Lenci, A., Sakti, S. and Xue, N., (eds.) 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 20-25 May 2024, Torino, Italy. ELRA and ICCL, pp. 6739-6751. ISBN: 978-2-493814-10-4. ISSN: 2951-2093.

Examining the Limitations of Computational Rumor Detection Models Trained on Static Datasets

Yida Mu, Xingyi Song, Kalina Bontcheva, Nikolaos Aletras
Department of Computer Science, The University of Sheffield
{y.mu, x.song, k.bontcheva, n.aletras}@sheffield.ac.uk

Abstract

A crucial aspect of a rumor detection model is its ability to generalize, particularly its ability to detect emerging, previously unknown rumors. Past research has indicated that content-based (i.e., using solely source post as input) rumor detection models tend to perform less effectively on unseen rumors. At the same time, the potential of context-based models remains largely untapped. The main contribution of this paper is in the in-depth evaluation of the performance gap between content and context-based models specifically on detecting new, unseen rumors. Our empirical findings demonstrate that context-based models are still overly dependent on the information derived from the rumors' source post and tend to overlook the significant role that contextual information can play. We also study the effect of data split strategies on classifier performance. Based on our experimental results, the paper also offers practical suggestions on how to minimize the effects of temporal concept drift in static datasets during the training of rumor detection methods.

Keywords: Rumor Detection, Computational Social Science, Computational Misinformation Analysis

1. Introduction

False rumors are claims or stories that are intended to deceive or mislead the public and can spread faster through social media, causing harm and confusion (Lazer et al., 2018; Zubiaga et al., 2018; Vosoughi et al., 2018). Due to their large volume and high velocity of spread, computational approaches (e.g., supervised rumor detection models) are typically employed to detect and analyze false rumors at an early stage (Bian et al., 2020; Lin et al., 2022; Tian et al., 2022).

Specifically, the task of rumor detection typically distinguishes the detection of check-worthy unverified claims (i.e., rumors) from other kinds of posts in social media (non-rumors) (Zubiaga et al., 2018). On the other hand, rumor verification[1] is typically the task of classifying a rumor as *True, False, Unverified, or Non-Rumor* (Kochkina et al., 2023).

Current computational rumor detection systems typically follow a two-step approach: (i) features are extracted from the textual content of the rumor (e.g., source post) along with contextual information,[2] and then (ii) models are trained and evaluated on static datasets using random data splits (Ma et al., 2016, 2017).

As demonstrated by Mu et al. (2023a); Hu et al. (2023), the evaluation of rumor detection systems performed on static datasets using random splits might not provide an accurate picture of the generalizability of such models to unseen rumors. Note that the evaluation conducted by Mu et al. (2023a); Hu et al. (2023) focuses solely on standard text classifiers (such as logistic regression) using only features derived from source posts.

However, rumors in social media also come with a rich amount of contextual information, including comments, user profile features and images, which complement the text of the source posts. For example, Figure 1 shows two Weibo users who post the same rumor about the death of a famous Chinese actor. Despite the source posts being identical, the remaining contextual information (e.g., comments and user profile attributes) is completely different. Note that the development of the majority of current rumor detection models relies on context-based features and utilizes random data splits (Bian et al., 2020; Rao et al., 2021).

The question that emerges is whether rumor detection models trained with contextual information using random data splits may also exhibit a tendency towards overestimation. Therefore, this paper primarily focuses on a systematic evaluation of the actual generalization capabilities (i.e., detecting rumors that are not previously known) of context-based rumor detection models, which is a hitherto unstudied research question.

The four contributions of this work are:

- Empirical proof (§ 4.1 & 5) that despite having additional contextual information, rumor detection models still struggle to detect unseen rumors appearing at a future date, with some models performing even worse than random

[1] In this work, for brevity, we refer to both tasks as rumor detection.

[2] In this work, we use the term 'contextual information' to refer to different forms of information associated with a rumor on social media, such as comments, images, and user profile attributes. The term 'content-based methods' refers to the use of only source posts as the model input.

LREC-COLING 2024, pages 6739–6751

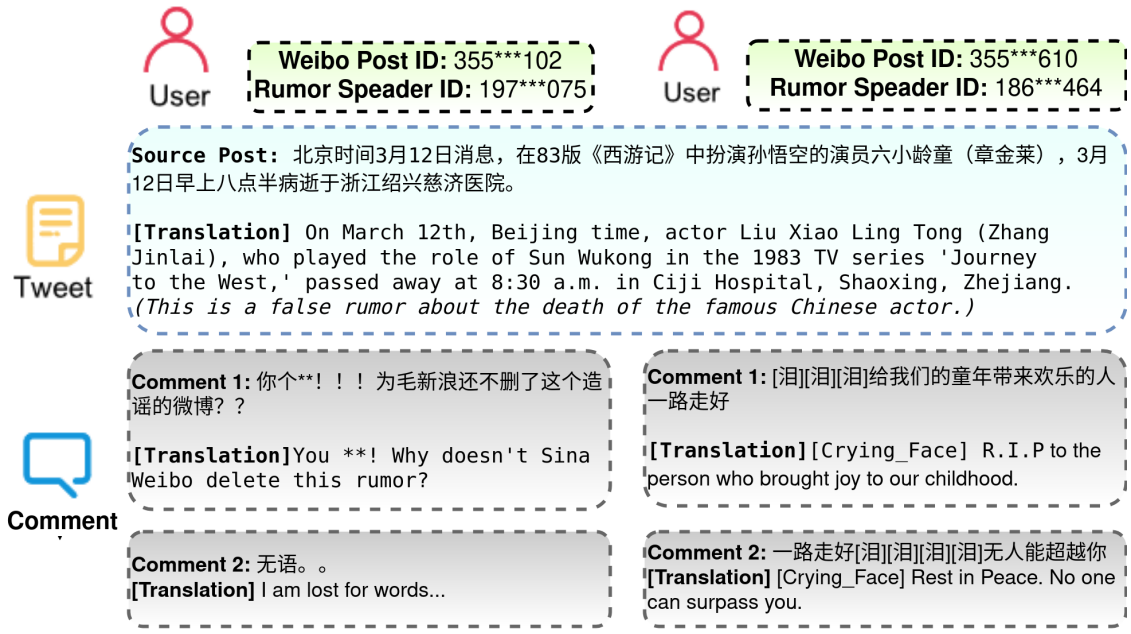

Figure 1: Two rumor spreaders (in the green box) posted an identical rumor and received different stances of comments (in the gray box), i.e., denial (on the left) and support (on the right), respectively. '[Crying_Face]' denotes the Loudly Crying Face emoji.

baselines (see Table 3).

- An ablation study (§ 5.3) that removes source posts from the inputs, revealed that current rumor detection approaches rely excessively on information from the source post, while neglecting the contextual information.
- A follow-up similarity analysis (§ 5.4) on content and context-based features, which elucidates the impact of training/test split strategies on model performance.
- Finally, we focus on the issue of effectively utilizing static datasets for rumor detection by providing practical recommendations (§ 6), such as implementing additional cleaning measures for the static dataset and enhancing the current evaluation metrics.

2. Related Work

2.1. Computational Rumor Detection Approaches

The increased consumption of news and information on social platforms has necessitated large-scale automated detection of unreliable content (Shu et al., 2017; Shearer and Gottfried, 2017), which led to the development of new rumor detection approaches based on state-of-the-art NLP techniques.

Early studies typically relied on handcrafted features extracted from source posts and user profile attributes using traditional machine learning models, such as SVM and Random Forest. (Qazvinian et al., 2011; Takahashi and Igata, 2012; Yang et al., 2012; Ma et al., 2015). With the emergence of neural-based NLP models (Mikolov et al., 2013), rumors started to be modeled with contextual embeddings such as Glove (Pennington et al., 2014) and ELMo (Peters et al., 2018). In addition, graph-based neural models have been employed to learn relationships from the propagation network of rumors, which includes retweet and comment chains (Bian et al., 2020; Lin et al., 2021; Yang et al., 2021). Other methods adopted multi-modal approaches to go beyond text and capture information from images (Wang et al., 2020; Sun et al., 2021; Zhou et al., 2022).

Recent hybrid models began including contextual information to improve rumor prediction performance (Lu and Li, 2020; Rao et al., 2021; Tian et al., 2022). The top-performing rumor detection systems (e.g., DUCK (Tian et al., 2022)) rely both on contextual information and user-level attributes, with 98 F1-measure on widely used datasets such as Weibo 16 (Ma et al., 2017) and CoAID (Cui and Lee, 2020).

Most of these rumor detection approaches however have a major weakness, as they are trained using random data splits which ignore a key temporal dimension of rumors and thus tend to overestimate model performance of future unseen rumors (Huang and Paul, 2018; Søgaard et al., 2021).

Statistic	Twitter 15	Twitter 16	Weibo 16	Weibo 20	Sun-MM
# of source posts	1,490	818	4,664	6,068	2374
# of True rumors	374	205	2,351	3,034	1,688
# of False rumors	370	205	2,313	3,034	686
# of Unverified rumors	374	203	-	-	-
# of Non-rumors	372	205	-	-	-
Average length of posts	19	19	105	88	-
Average # of comments	22	16	804	62	-
Average length of comments	242	202	8,484	13,592	-
Contextual Information					
Source Posts	✓	✓	✓	✓	✓
Comments	G	G	G+S	S	-
User Profile Attributes	✓	✓	✓	✓	✓
Images	-	-	-	-	✓

Table 1: Dataset statistics. 'G' and 'S' denote comment propagation network (Graph) and comment sequence (S) respectively. We also present contextual-based features obtained from each dataset.

2.2. The Effect of Temporal Concept Drift in NLP Downstream Tasks

Previous work on legal, abusive language, COVID-19, and biomedical classification tasks (Huang and Paul, 2019, 2018; Chalkidis and Søgaard, 2022; Mu et al., 2023b; Jin et al., 2023) has investigated the sensitivity of classifiers to temporal concept drift (i.e., the deterioration of their performance due to temporal/topic variation) when evaluated on chronological data splits. However, temporal concept drift mainly affects the rumor text (i.e. new unseen topics), as rumors on the same topic posted by different users have different contextual information. Mu et al. (2023a) explore the impact of temporal concept drift on rumor detection using standard text classifiers such as logistic regression and fully fine-tuned BERT.

In contrast, this paper performs an extensive empirical evaluation of the effect of temporal concept drift on neural rumor detection models which combine textual and contextual information.

3. Experimental Setup

3.1. Data

For comprehensiveness and reliability, our experiments are carried out on five datasets (see Table 1 for details), which have been widely used in prior rumor detection research (Bian et al., 2020; Rao et al., 2021; Sun et al., 2021; Tian et al., 2022; Lin et al., 2022):

- **Twitter 15 & Twitter 16** (Ma et al., 2017) are two English datasets that include tweets categorized into one of four categories: *True Rumor (T), False Rumor (F), Non-rumor (NR) and Unverified Rumor (U)*.
- **Weibo 16** (Ma et al., 2017) consists of 4,664 Weibo posts in Chinese. It comprises 2,313 *false rumors* debunked by the official Weibo Fact-checking Platform and 2,351 *non-rumors* sourced from mainstream news sources.
- **Weibo 20** (Rao et al., 2021) is a Chinese rumor detection dataset similar to Weibo 16. It provides 3,034 *non-rumors* and 3,034 *false rumors* from the same Weibo fact-checking platform as Weibo 16.
- **Sun-MM** (Sun et al., 2021) comprises 2,374 annotated tweets (i.e., *rumor or non-rumor*) that cover both textual (i.e., source post) and visual (i.e., image) information. It is typically used for multi-modal rumor detection.

It should be noted that most prior rumor detection models are typically evaluated on two or three datasets only, typically from a specific language.

3.2. Models

Following (Kochkina et al., 2023), we evaluate a number of top-performing rumor detection models.[3]. Each dataset is used to train at least three models, based on the information it provides (see Table 2 for details).

Weak Baseline For reference, we provide a weak baseline by randomly generating predictions compared to the ground truth labels of the test set.

SVM-HF (Source Post + User Profile) Similar to (Yang et al., 2012; Ma et al., 2015), we use a

[3]Here, we only consider reproducible models with publicly available code and full implementation details. Note that these models have been extensively employed as baselines in prior research (Rao et al., 2021; Tian et al., 2022)

Models	Contextual Information				Datasets				
	Post	Comment	User	Image	Twitter 15	Twitter 16	Weibo 16	Weibo 20	Sun-MM
SVM-HF	✓	-	✓	-	✓	✓	✓	✓	✓
BERT	✓	-	-	-	✓	✓	✓	✓	✓
H-Trans	✓	✓	-	-	-	-	✓	✓	-
Bi-GCN	✓	✓	-	-	✓	✓	✓	-	-
Hybrid	✓	-	-	✓	-	-	-	-	✓

Table 2: Model details.

linear SVM model using source posts represented with TF-IDF and various handcrafted features extracted from user profile attributes e.g., number of followers, account status (i.e., whether a verified account or not), number of historical posts, etc.

BERT (Source Post) In line with previous work (Rao et al., 2021; Tian et al., 2022), we use solely source posts as input to fine-tune the Bert-base model[4] (Devlin et al., 2019) by adding a linear layer on top of the 12-layer transformer architecture with a softmax activation. We consider the special token '[CLS]' as the post-level representation.

Bi-GCN (Comment Network) To model the network of comment propagation, we use Bi-Directional Graph Convolutional Networks (Bi-GCN) (Bian et al., 2020). Bi-GCN employs two separate GCNs with (i) a top-down directed graph representing rumor spread to learn the patterns of rumor propagation; and (ii) another GCN with an opposite directed graph of rumor diffusion.

Hierarchical Transformers (Source Post + Comment Sequence) Similar to prior work (Rao et al., 2021; Tian et al., 2022), we use a hierarchical transformer-based network to encode separately the source post and its sequence of comments.[5] We then add a self-attention and a linear projection layer with softmax activation to combine the hidden representation of posts and comments.

Hybrid Vision-and-Language Representation (Source Post + Image) We use visual transformer[6] (ViT) (Dosovitskiy et al., 2020) and BERT (Devlin et al., 2019) to represent images and source posts of rumors for the Sun-MM dataset. We then combine the two hidden representations by adding a fully connected layer with softmax activation for rumor classification.

3.3. Data Pre-processing

We begin by processing all the source posts and comments, replacing @mentions and links with special tokens such as '@USR' and 'URL' respectively. For the English datasets, we also convert all tweets to lowercase before feeding them to the bert-base-uncased model.

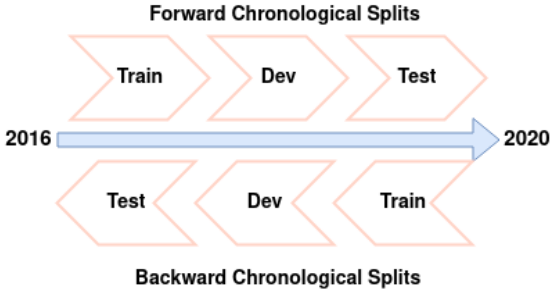

Figure 2: An example of using forward and backward chronological data splits on Weibo 20 dataset (including rumors from 2016 to 2020). There is no overlap among the three subsets.

3.4. Evaluation Metrics

We run each model three times with different random seeds. In accordance with the original settings (Ma et al., 2016, 2017; Rao et al., 2021), we report the average macro precision, recall, F1-score, and accuracy for all binary datasets, i.e., Weibo 16, Weibo 20, and Sun-MM. Since the Twitter datasets (Twitter 15 & 16) have multi-class labels, we report the average accuracy and F1-score for each class.

3.5. Hyper-parameters

For linear SVM, we use word-level and character-level tokenizers for English and Chinese datasets respectively. We set learn rate as 2e-5 and batch size as 16 for the Bert-base model. For all transformer-based models, we set the max input length as 512 covering all posts. The implementation details of Bi-GCN are available from the open-source repositories.[7] All experiments are performed using a single Nvidia RTX Titan GPU with 24GB memory.

[4]We use bert-base-uncased and bert-base-chinese models from Hugging Face (Wolf et al., 2020) for English and Chinese datasets respectively.

[5]Given that the total number of tokens of the source post and all comments exceeds the maximum input length (i.e., 512 tokens) of most Bert-style models.

[6]https://huggingface.co/google/vit-base-patch16-224

[7]https://github.com/TianBian95/BiGCN

Models & Splits		Twitter 15					Twitter 16				
		Acc.	*NR F1*	*F F1*	*T F1*	*U F1*	*Acc.*	*NR* F1	*F* F1	*T* F1	*U* F1
Weak Baseline		0.240	0.224	0.246	0.238	0.254	0.248	0.174	0.250	0.300	0.264
SVM-HF	*Random*	**0.739**	**0.727**	**0.701**	**0.803**	**0.728**	**0.709**	**0.697**	**0.602**	**0.858**	**0.663**
	Forward	0.413	0.589	0.366	0.092	0.304	0.373	0.523	0.226	0.297	0.214
	Reverse	0.353	0.590	0.462	0.063	0.062	0.380	0.520	0.103	0.411	0.368
BERT	*Random*	**0.615**	**0.561**	**0.593**	**0.692**	**0.599**	**0.598**	0.381	**0.615**	**0.698**	**0.625**
	Forward	0.366	0.382	0.226	0.457	0.328	0.380	**0.446**	0.306	0.110	0.489
	Reverse	0.367	0.430	0.256	0.455	0.292	0.428	0.371	0.210	0.662	0.483
Bi-GCN	*Random*	**0.838**	**0.785**	**0.841**	**0.886**	**0.785**	**0.854**	**0.745**	**0.861**	**0.939**	**0.847**
	Forward	0.415	0.509	0.386	0.311	0.319	0.489	0.551	0.381	0.401	0.511
	Reverse	0.498	0.584	0.339	0.786	0.118	0.517	0.502	0.413	0.667	0.419

Table 3: Experimental results of Twitter 15 & 16 datasets across three different data split strategies. Cells in **bold** indicate the best results from all models. Cells in gray indicate that the model trained using random splits achieves significantly better performance than using both forward and backward chronological splits. ($p < 0.05$, t-test).

Models	Splits	*Weibo 16*				*Weibo 20*				Sun-MM			
		Acc.	*P*	*R*	*F1*	*Acc.*	*P*	*R*	*F1*	*Acc.*	*P*	*R*	*F1*
Weak Baseline		0.493	0.493	0.492	0.493	0.501	0.501	0.501	0.501	0.514	0.512	0.514	0.512
SVM-HF	*Random*	**0.906**	**0.907**	**0.906**	**0.906**	**0.870**	**0.870**	**0.868**	**0.870**	**0.783**	**0.742**	**0.758**	**0.749**
	Forward	0.823	0.855	0.822	0.819	0.680	0.691	0.680	0.676	0.689	0.635	0.630	0.635
	Backward	0.752	0.757	0.752	0.752	0.801	0.802	0.801	0.801	0.771	0.740	0.676	0.692
BERT	*Random*	**0.918**	**0.918**	**0.917**	**0.918**	**0.920**	**0.921**	**0.920**	**0.920**	**0.839**	**0.807**	**0.806**	**0.806**
	Forward	0.889	0.892	0.888	0.888	0.738	0.756	0.738	0.732	0.708	0.682	0.708	0.680
	Backward	0.809	0.812	0.809	0.808	0.898	0.899	0.898	0.898	0.807	0.783	0.735	0.748
Bi-GCN	*Random*	**0.892**	**0.893**	**0.885**	**0.887**	**-**	**-**	**-**	**-**	-	-	-	-
	Forward	0.843	0.843	0.834	0.835	**-**	**-**	**-**	**-**	-	-	-	-
	Backward	0.762	0.783	0.762	0.747	**-**	**-**	**-**	**-**	-	-	-	-
H-Trans / Hybrid	*Random*	**0.955**	**0.956**	**0.955**	**0.955**	**0.959**	**0.960**	**0.959**	**0.959**	**0.853**	**0.818**	**0.829**	**0.823**
	Forward	0.946	0.949	0.946	0.946	0.850	0.860	0.849	0.850	0.707	0.687	0.725	0.685
	Backward	0.792	0.833	0.785	0.793	0.940	0.938	0.935	0.938	0.821	0.782	0.805	0.791

Table 4: Experimental results of Weibo 16 & 20 and Sun-MM across three different data split strategies. Cells in **bold** indicate the best results from all models. Cells in gray indicate that the model trained using random splits achieves significantly better performance than using both forward and backward chronological splits. ($p < 0.05$, t-test).

4. Evaluation Strategies

4.1. Data Splits

To examine the effect of data splitting strategies on the models' predictive performance, we compare three strategies: the widely used random data split against two types of chronological data splits (see Figure 2).

- **Forward Chronological Splits** For each dataset, we initially sort all rumors chronologically, from the oldest to the newest. We then divide them into three subsets: a training set (containing 70% of the oldest rumors), a development set (10% of the rumors that were posted after those in the training set but before those in the test set), and a test set (containing the 20% most recent rumors). This data split strategy allows the model to be trained and fine-tuned on *older rumors* and then be evaluated on the most *recent ones*.
- **Backward Chronological Splits** In contrast, here all rumors are sorted starting from the most recent ones to the oldest ones, and then are split in the same way as the forward chronological splits. This allows the model to be trained on the *newest rumors* and evaluated on the *oldest ones*.
- **Random Splits** This is the most commonly adopted data split strategy in prior work. All datasets are divided into three subsets using a stratified random split approach[8].

These two different temporal split strategies enable the evaluation of temporal concept drift effects on model performance.

Some prior rumor detection research has used a leave-a-rumor-out strategy (Lukasik et al., 2015, 2016), where each dataset is divided into N folds, where N denotes the the number of unique rumor

[8]We use a data split tool from sklearn: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

events in the given dataset. In this case, rumor detection models are evaluated using N-fold cross validation, i.e., using $N-1$ unique rumors and all associated posts as the training set and the posts about the last remaining rumor as the test set. In this way, it is possible to evaluate model performance on *new unseen rumors*. However, it has not been possible to experiment with this data split protocol as none of the datases used in this paper cluster posts into individual events which give rise to a unique rumor, with associated multiple social media posts about it.

5. Results and Discussion

5.1. Model Performance on Random Splits

The experimental results for all rumor detection approaches and data split strategies are shown in Tables 3 and 4. We can observe that training on random splits always leads to significant overestimation (t-test, $p < 0.01$) of model accuracy as compared to training on both forward and backward chronological splits.

Taking the best performing Bi-GCN model on Twitter 15 as an example, we observe a decrease in model accuracy of at least 39.4% when comparing test results on random splits against the two chronological splits. Furthermore, we find that some models (e.g., SVM-HF and Bi-GCN on Twitter 15) perform even worse than a weak baseline (e.g., the F1-measure results for the false rumor category (F) across two chronological splits in comparison with the weak baseline) that uses random predictions. As expected, our empirical findings align with previous studies of temporal impact in other downstream NLP tasks (Huang and Paul, 2019; Chalkidis and Søgaard, 2022; Mu et al., 2023b).

The results indicate that models learn to classify accurately rumor posts in the test set only when they are highly similar to posts in the training data, even though the remaining contextual information (such as user profile attributes, comments, and sometimes images) are different. To further investigate the impact of this semantic overlap, we conduct an ablation study (Section 5.3) and a similarity analysis (Section 5.4).

5.2. Forward v.s. Backward Chronological Splits

Our experimental results show that models trained using backward chronological splits achieve higher accuracy on all datasets (except Weibo 16) as compared to those on forward chronological splits. This suggests that the models have the tendency to learn recurrent rumors. This observation is consistent across datasets. For instance, the accuracy of all models on the Twitter 16 dataset is higher when random splits are used for training as compared to forward splits, but lower when compared to backward splits. This may be attributed to similarities between the training and test sets. This is investigated further in Section 5.4.

5.3. Ablation Study

In order to evaluate the impact of the source post's text on rumor detection performance, we perform a source post removal ablation study[9]. Our hypothesis is that after removing the source posts, there will be no significant difference in the performance of the rumor detection models trained according to the different data split strategies. We conduct experiments using (i) SVM-HF on all datasets, (ii) the Hier-Transformer model on Weibo 16 and Weibo 20, and (iii) visual transformer (ViT) on Sun-MM dataset.

The results of the ablation study are reported in Table 6 and Table 5. We demonstrate that when the source posts are removed from the input, all models except for ViT model (see Section 5.4 for further analysis) no longer exhibit consistent superiority over forward and backward chronological splits as compared to using random splits. As we have shown, two identical rumors can have different contextual information. This indicates that temporalities are not commonly reflected in the majority of contextual information associated with rumors in social media. Notably, even without the source post, the H-Trans model can achieve competitive performance using chronological splits. For instance, it achieves up to 93.8% and 94.4% accuracy on Weibo 16 and Weibo 20, respectively, which is comparable to the performance of the Bi-GCN and original H-Trans models (which take the source post as input). We hypothesize that rumor debunking information may be present in the comments (for example, see Figure 1), which can assist in the decision-making process of the rumor classifier. Next we conduct linguistic analysis to elucidate the distinctions between comments from rumors and non-rumors in Weibo 16 & 20.

5.4. Similarity Analysis

This section explores the impact of data split strategies on the content and contextual information in the respective training and test sets. We investigate whether a decrease in model predictive per-

[9]Previous ablation studies have focused primarily on removing new features rather than source posts (Sun et al., 2021; Tian et al., 2022)

Models	Splits	Twitter 15					Twitter 16				
		Acc.	*NR*	*F*	*T*	U	*Acc.*	*NR*	*F*	*T*	U
SVM-HF w/o SP	*Random*	**0.383**	0.609	0.050	0.356	**0.132**	0.343	0.494	0.140	**0.273**	**0.229**
	Forward	0.375	**0.635**	0.039	**0.374**	0.086	**0.417**	**0.689**	**0.333**	0.046	0.158
	Reverse	0.361	0.590	**0.133**	0.359	0.050	0.328	0.499	0.178	0.170	0.021

Table 5: Ablation study of Twitter 15 & 16 datasets across three different data split strategies. Cells in **bold** indicate the best results from all models.

Models	Splits	Weibo 16				Weibo 20				Sun-MM			
		Acc.	*P*	*R*	*F1*	*Acc.*	*P*	*R*	*F1*	*Acc.*	*P*	*R*	*F1*
SVM-TS w/o SP	*Random*	0.887	0.889	0.887	0.887	0.773	0.801	0.773	0.768	0.707	**0.663**	0.510	0.439
	Forward	**0.936**	**0.944**	**0.936**	**0.936**	0.699	0.753	0.699	0.681	0.701	0.602	0.507	0.434
	Reverse	0.683	0.698	0.684	0.678	**0.831**	**0.837**	**0.831**	**0.830**	**0.713**	0.655	**0.52**	**0.453**
H-Trans / Hybrid w/o SP	*Random*	0.929	0.930	0.929	0.929	0.925	0.926	0.925	0.925	**0.726**	**0.674**	**0.691**	**0.681**
	Forward	**0.938**	**0.935**	**0.934**	**0.935**	0.851	0.856	0.851	0.851	0.655	0.521	0.514	0.505
	Reverse	0.730	0.795	0.732	0.715	**0.944**	**0.945**	**0.944**	**0.944**	0.623	0.516	0.514	0.514

Table 6: Ablation study of Weibo 16 & 20 and Sun-MM datasets across three different data split strategies. Cells in **bold** indicate the best results from all models.

Dataset	Splits	IOU	DICE	Acc.
Twitter 15	*Random*	19.6	23.2	0.615
	Forward	11.2	13.8	0.366
	Backward	11.7	14.5	0.367
Twitter 16	*Random*	17.1	20.3	0.598
	Forward	9.9	12.3	0.380
	Backward	10.6	13.1	0.428
Weibo 16 Source Post	*Random*	28.4	32.8	0.918
	Forward	23.5	28.4	0.892
	Backward	22.3	27.2	0.812
Weibo 20 Source Post	*Random*	26.2	30.2	0.920
	Forward	20.9	24.6	0.738
	Backward	21.8	26.2	0.898
Sun-MM	*Random*	23.5	27.2	0.839
	Forward	14.0	16.6	0.708
	Backward	13.4	17.3	0.807
Weibo 16 Comment	*Random*	26.7	31.3	0.929
	Forward	26.2	31.2	0.938
	Backward	25.4	30.0	0.730
Weibo 20 Comment	*Random*	26.0	30.3	0.925
	Forward	25.5	30.1	0.851
	Backward	24.8	28.5	0.944

Table 7: Textual similarity betweet training and test sets using random and temporal data splits.

formance occurs due to variations between the two subsets used for training and testing, and whether the difference in performance lessens as the datasets become more similar to each other.

Source Post Similar to Kochkina et al. (2023); Mu et al. (2023b), we first measure the difference in textual similarity between training and test sets generated using random and chronological data splits using two standard matrices with ranges from 0 to 1.

Intersection over Union (IoU) (Tanimoto, 1958)

$$IoU = \frac{|V^{Train} \cap V^{Test}|}{|V^{Train} \cup V^{Test}|} \quad (1)$$

DICE coefficient (DICE) (Dice, 1945)

$$DICE = \frac{2 \times |V^{Train} \cap V^{Test}|}{|V^{Train}| + |V^{Test}|} \quad (2)$$

where V^{Train} and V^{Test} refer to the set of unique words from training and test sets; and $|V^{Train} \cap V^{Test}|$ and $|V^{Train} \cup V^{Test}|$ indicate the number of unique words that appear in the *intersection* and *union* of training and test sets respectively. When the two sets have no shared vocabulary list, the IoU and DICE values will be 0, while if they are identical, the IoU and DICE values will be 1.

We display the similarity of the source posts between training and test sets using different data split strategies in Table 7. Additionally, we provide the accuracy of the BERT model (which takes only the source post as input) for each dataset as a reference.

We demonstrate that using random splits leads to significantly higher IoU and DICE values (t-test, $p < 0.001$), indicating greater similarities between the training and test sets compared to both forward and backward chronological splits. This suggests that rumors with similar content, resulting from temporal concept drift, appear in both training and test sets when employing random data splits. Additionally, we discover a positive correlation (using Pearson's Test) between model accuracy and the similarity distance of training and test sets, as measured by both IOU (Pearsons' r = 0.865, $p < 0.05$) and DICE (Pearsons' r = 0.879, $p < 0.001$) values. In other words, higher textual similarities correspond to better classifier performance.

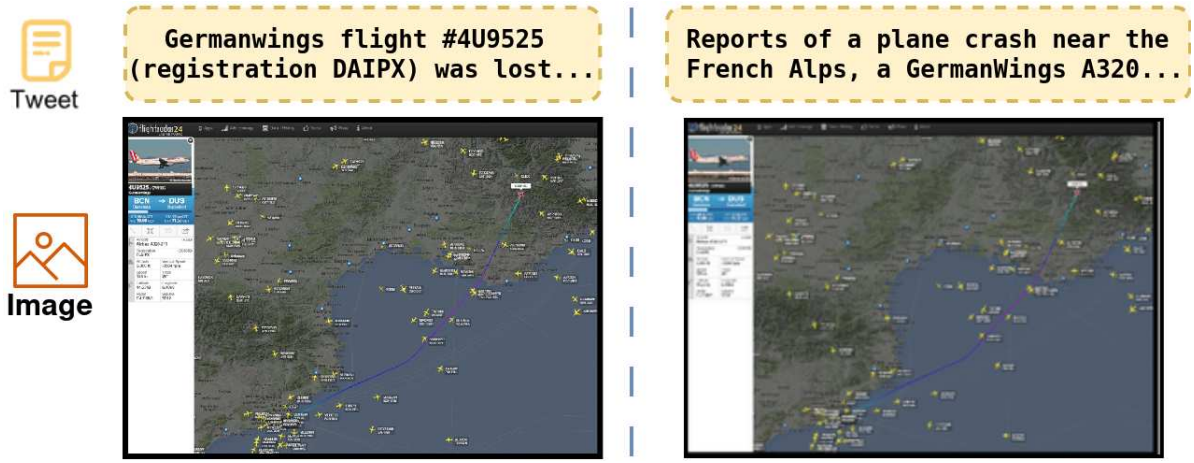

Figure 3: Two rumors from the Sun-MM Dataset related to the 'plane crash' event contain similar images and were published during a comparable time period. More examples are displayed in the Appendix (see Figure 4).

User Profile Attributes We use cosine similarity to assess the difference between the mean values of user profile attributes from the training and test sets. However, we do not observe a significant difference in cosine similarity values when using both random and chronological data splits, as all rumor speaders are unique across all datasets.

Comments Considering the comparable model performance, with accuracy of up to 93.8% and 94.4% on Weibo 16 and Weibo 20, when using only comments as input, we hypothesize that the comments from the two classes are significantly different. To identify the difference in comments that distinguish between rumors and non-rumors in Weibo 16 and Weibo 20, we employ the univariate Pearson's correlation test (Schwartz et al., 2013). We observe that there is a large amount of words related to debunking rumors (e.g., 'false', 'really?', and 'truth') in the comments associated with false rumors on both Weibo 16 and Weibo 20. On the other hand, comments associated with non-rumors are more words related to the daily life of the public. Note that non-rumors in Weibo datasets are collected from mainstream media accounts.

Images Ablation study results (see Table 6) show that only the ViT model, which uses images alone as input, is affected by the temporal data splits (i.e., the deterioration of model performance). We further explore the Sun-MM dataset and uncover that rumors with similar content are usually posted with similar images. We show examples in Figure 3 . Note that similar semantic objects (e.g., entitles (Sun et al., 2021)) can be extracted from similar images, which can impact the accuracy of the model.

6. How do we properly use static datasets?

Apart from prioritizing skewed methodologies solely for achieving high accuracy on rumor detection datasets, it is essential to develop a deeper comprehension of the protocol we employ and generate significant insights. Given the limitations raised by our experiments, we make the following practical suggestions for developing new rumor detection systems on static datasets:

- For practical applications that aim to detect **unseen rumors**, it is essential to consider chronological splits when evaluating all rumor detection approaches on static datasets, in addition to standard random splits. By using forward and backward chronological splits, we can assess the ability of the rumor classifiers to handle both earlier and older unseen rumors.

- Considering that temporalities (i.e., the temporal concentration of rumor topics) typically occur in widely used rumor detection datasets (e.g., Twitter 15&16 and Weibo 16 (Ma et al., 2016, 2017)), one can apply an additional data pre-processing measure to filter out rumor events with multiple posts. For instance, using out-of-the-box methods such as Levenshtein distance (Levenshtein et al., 1966) and BERTopic (Grootendorst, 2022), we identified a total of 9 similar rumors that resemble the false rumor depicted in Figure 1. After conducting a more in-depth error analysis on the predictions generated by the H-Trans model, which has demonstrated the highest predictive performance on Weibo 16, we discovered that the models can accurately classify all of

these rumors in the test set when employing random data splits.

- Current evaluation metrics, such as accuracy and F1-measure, are unable to accurately assess the true capability of rumor classifiers in detecting unseen rumors. Therefore, there is a need for new measures to evaluate the accuracy of model predictions for unknown rumors. For example, one can calculate the accuracy of a rumor detection system by excluding known rumors (i.e., similar rumors appearing in the training set) from the test set.

- Given the limitations of the current pipeline that relies solely on static datasets, we argue that evaluation models should not be restricted to such datasets. By leveraging the consistent format of datasets collected from the same platform (as shown in Table 1), for example, one can explore **broader temporalities** by training a rumor classifier on Twitter 15 and evaluating its performance on Twitter 16. This protocol enables a more comprehensive examination of the generalizability of rumor detection systems, which is crucial for their practical applications in the real world (Moore and Rayson, 2018; Yin and Zubiaga, 2021; Kochkina et al., 2023).

7. Conclusion

In this paper, we evaluate the limitations of existing widely used rumor detection models trained on static datasets. Through empirical analysis, we demonstrate that the use of chronological splits significantly diminishes the predictive performance of widely-used rumor detection models. To better understand the causes behind these limitations, we conduct a fine-grained similarity analysis and an ablations study. Finally, we provide practical recommendations for future research in the advancement of new rumor detection systems.

Limitations and Future Work We conducted an empirical study on current rumor detection models, utilizing both the source post and **standard contextual information** (such as comments, images, and user profile attributes) as input. However, previous research has employed hidden features, such as sentiment and entities, which can be extracted from the source post and contextual information (Rao et al., 2021; Sun et al., 2021). We consider this as future work and aim to explore additional feature settings. Besides, the current work is limited to English and Chinese, and we acknowledge that further research into more multilingual datasets should be considered in the future.

Ethics Statement

Our work has been approved by the Research Ethics Committee of the University of Sheffield, and complies with the data policies of Twitter[10] and Weibo[11]. All datasets are obtained through the links provided in the source papers.

Acknowledgments

This research is supported by an EU Horizon 2020 grant (agreement no.871042) ("So-BigData++: European Integrated Infrastructure for Social Mining and BigData Analytics").[12]

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 549–556.

Ilias Chalkidis and Anders Søgaard. 2022. Improved multi-label classification under temporal concept drift: Rethinking group-robust algorithms in a label-wise setting. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2441–2454, Dublin, Ireland. Association for Computational Linguistics.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

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.

Lee R Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai,

[10] https://developer.twitter.com/en/docs/twitter-api

[11] https://open.weibo.com

[12] http://www.sobigdata.eu

Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Zhengjia Wang, and Zhiwei Jin. 2023. Learn over past, evolve for future: Forecasting temporal trends for fake news detection. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics: Industry Track*. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Mali Jin, Yida Mu, Diana Maynard, and Kalina Bontcheva. 2023. Examining temporal bias in abusive language detection. *arXiv preprint arXiv:2309.14146*.

Elena Kochkina, Tamanna Hossain, Robert L Logan IV, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, 60(1):103116.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, pages 707–710. Soviet Union.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556.

Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor detection on Twitter with claim-guided hierarchical graph attention networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10035–10047, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. 2015. Classifying tweet level judgements of rumours in social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2590–2595, Lisbon, Portugal. Association for Computational Linguistics.

Michal Lukasik, P. K. Srijith, Duy Vu, Kalina Bontcheva, Arkaitz Zubiaga, and Trevor Cohn. 2016. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 393–398, Berlin, Germany. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 3818–3824.

Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 1751–1754.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Andrew Moore and Paul Rayson. 2018. Bringing replication and reproduction together with generalisability in nlp: Three reproduction studies for target dependent sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1132–1144.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023a. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 724–731.

Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2023b. Examining temporalities on stance detection towards covid-19 vaccination. *arXiv preprint arXiv:2304.04806*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1589–1599.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. 2021. STANKER: Stacking network based on level-grained attention-masked BERT for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3347–3363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Elisa Shearer and Jeffrey Gottfried. 2017. News use across social media platforms 2017.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Mengzhu Sun, Xi Zhang, Jianqiang Ma, and Yazheng Liu. 2021. Inconsistency matters: A knowledge-guided dual-inconsistency network for multi-modal rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1412–1423, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tetsuro Takahashi and Nobuyuki Igata. 2012. Rumor detection on twitter. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 452–457. IEEE.

Taffee T Tanimoto. 1958. Elementary mathematical theory of classification and prediction.

Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2022. DUCK: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949, Seattle, United States. Association for Computational Linguistics.

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

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake

news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 540–547.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7.

Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2021. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 1417–1423.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Honghao Zhou, Tinghuai Ma, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2022. Mdmn: Multi-task and domain adaptation based multi-modal network for early rumor detection. *Expert Systems with Applications*, 195:116517.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys*, 51(2):1–36.

Appendix

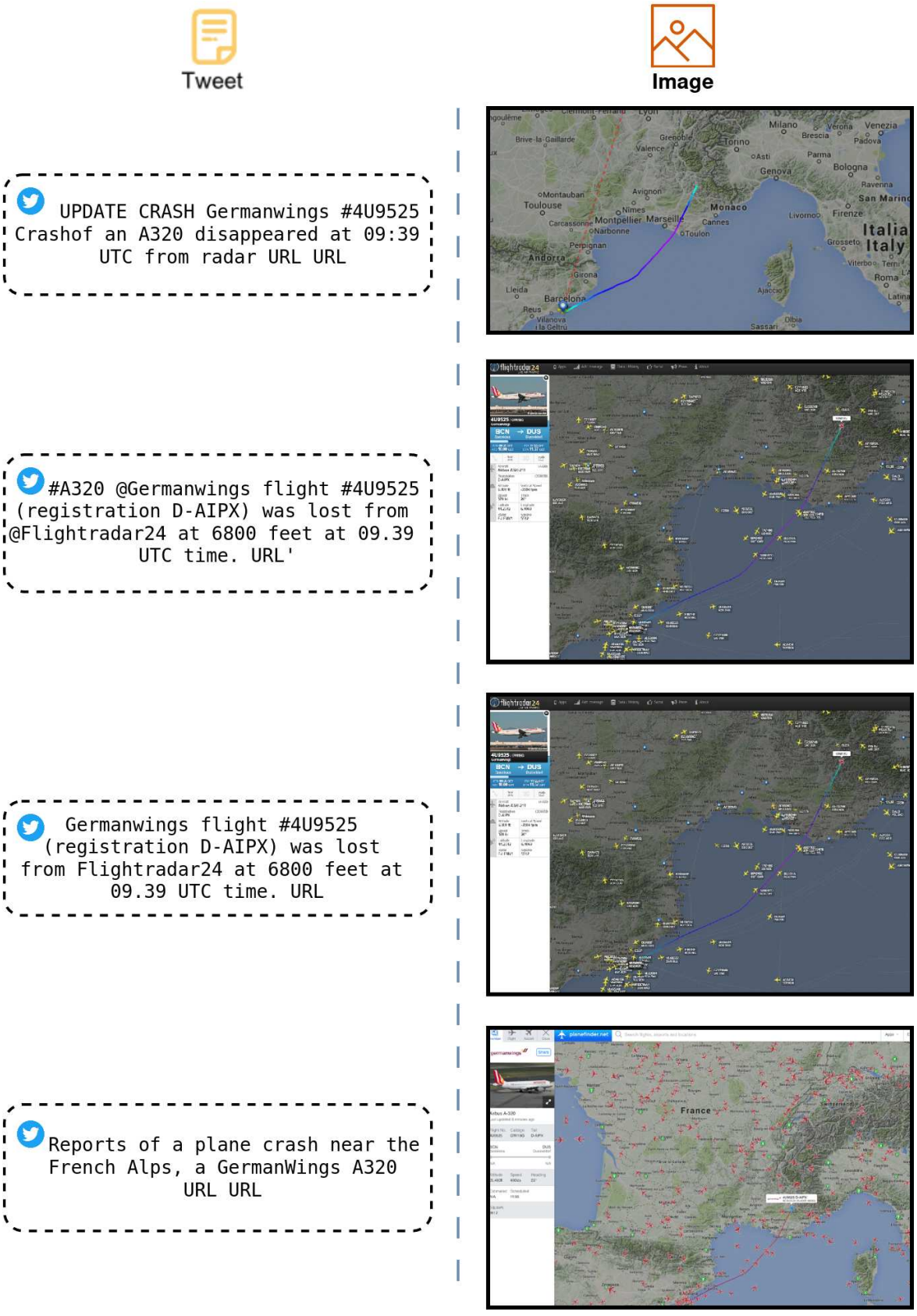

Figure 4: Four pairs of rumors related to the 'plane crash' event (from the Sun-MM Dataset) contain similar images and were published during a comparable time period.