Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection

Beizhe Hu[1,2] **Qiang Sheng**[1,2,*] **Juan Cao**[1,2] **Yongchun Zhu**[1,2]
Danding Wang[1] **Zhengjia Wang**[1,2] **Zhiwei Jin**[3]

[1]Key Lab of Intelligent Information Processing of Chinese Academy of Sciences,
Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3] ZhongKeRuijian Technology Co., Ltd.
{hubeizhe21s,shengqiang18z,caojuan,zhuyongchun18s}@ict.ac.cn
{wangdanding,wangzhengjia21b}@ict.ac.cn, jinzhiwei@ruijianai.com

Abstract

Fake news detection has been a critical task for maintaining the health of the online news ecosystem. However, very few existing works consider the temporal shift issue caused by the rapidly-evolving nature of news data in practice, resulting in significant performance degradation when training on past data and testing on future data. In this paper, we observe that the appearances of news events on the same topic may display discernible patterns over time, and posit that such patterns can assist in selecting training instances that could make the model adapt better to future data. Specifically, we design an effective framework **FTT** (**F**orecasting **T**emporal **T**rends), which could forecast the temporal distribution patterns of news data and then guide the detector to fast adapt to future distribution. Experiments on the real-world temporally split dataset demonstrate the superiority of our proposed framework. The code is available at https://github.com/ICTMCG/FTT-ACL23.

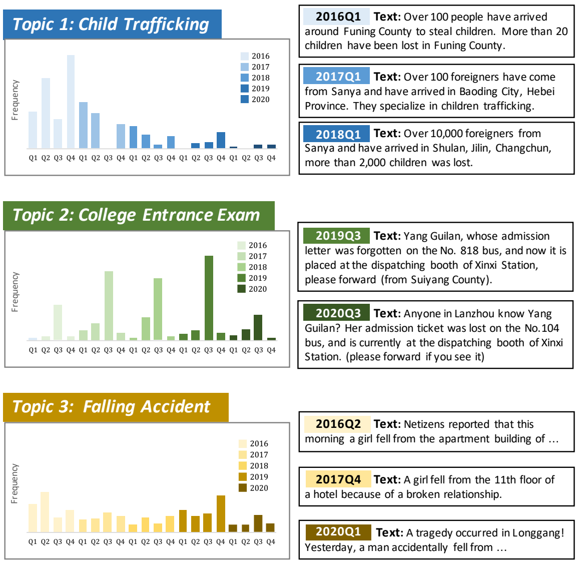

Figure 1: Topic-level statistics of news items across five years in our data. We see that different topics present diverse temporal patterns such as *decrease* (Topic 1), *periodicity* (Topic 2), and *approximate stationery* (Topic 3), which we rely on to forecast temporal trends for better fake news detection in the future. The case texts are translated from Chinese into English.

1 Introduction

Automatic fake news detection, which aims at distinguishing inaccurate and intentionally misleading news items from others automatically, has been a critical task for maintaining the health of the online news ecosystem (Shu et al., 2017). As a complement to manual verification, automatic fake news detection enables efficient filtering of fake news items from a vast news pool. Such a technique has been employed by social media platforms like Twitter to remove COVID-19-related misleading information during the pandemic (Roth, 2022).

Over the past decade, most fake news detection researchers have followed a conventional paradigm of collecting a fixed dataset and *randomly* dividing it into training and testing sets. However, the assumption that news data subsets are independent and identically distributed often does not hold true in real-world scenarios. In practice, a fake news detection model is trained on *offline data* collected up until the current time period but is required to detect fake news in newly arrived *online data* at the upcoming time period. Due to the rapidly-evolving nature of news, news distribution can vary with time, namely *temporal shift* (Du et al., 2021; Gaspers et al., 2022), leading to the distributional difference between offline and online data. Recent empirical studies (Zhang et al., 2021; Mu et al., 2023) evidence that fake news detection models suffer significant performance degradation when the dataset is temporally split. Therefore, the temporal shift issue has been a crucial obstacle to real-world fake news detection systems.

The temporal shift scenario presents a more significant challenge than common *domain shift* sce-

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 116–125
July 10-12, 2023 ©2023 Association for Computational Linguistics

narios. Most existing works on the domain shift in fake news detection focus on transfer among predefined news channels (e.g., politics) (Silva et al., 2021b; Mosallanezhad et al., 2022; Lin et al., 2022; Nan et al., 2022). However, consecutive data slices over time have various types of temporal dependencies and non-explicit distributional boundaries, making the temporal shift challenging. Moreover, these works assume the availability of target domain data, which is impossible for the temporal shift scenarios. Under such constraints, our aim is to train a model using presently available data to generalize to future online data (corresponding to temporal generalization task; Wang et al., 2022). Others that improve the generalizability to unseen domains learn domain-invariant features by adversarial learning (Wang et al., 2018) and domain-specific causal effect removal (Zhu et al., 2022a), but do not consider the characteristics of temporal patterns of news events.

In this paper, we posit that the appearance of news events on the same topic presents diverse temporal patterns, which can assist in evaluating the importance of previous news items and boost the detection of fake news in the upcoming time period. In Figure 1, we exemplify this assumption using the statistics of news items on three topics in the Chinese Weibo dataset: Topic 1 presents the temporal pattern of *decrease*, where news about child trafficking becomes less frequent. Topic 2 presents the *periodicity* of news related to the college entrance exam which takes place annually in the second quarter (Q2).[1] In Topic 3, news items about falling accidents appear repeatedly and exhibit an *approximate stationary* pattern. Such temporal patterns indicate the different importance of news samples in the training set for detection in future quarters. For instance, instances of Topic 2 in the training set are particularly important for effectively training the detector to identify fake news in Q3.

To this end, we propose to model the temporal distribution patterns and forecast the topic-wise distribution in the upcoming time period for better temporal generalization in fake news detection, where the forecasted result guides the detector to fast adapt to future distribution. Figure 2 illustrates our framework **FTT** (**F**orecasting **T**emporal **T**rends). We first map training data to vector space and perform clustering to discover topics. Then we model the temporal distribution and forecast the frequency of news items for each topic using a decomposable time series model. Based on the forecasts, we evaluate the importance of each item in the training data for the next time period by manipulating its weight in training loss. Our contributions are summarized as follows:

- **Problem:** To the best of our knowledge, we are the first to incorporate the characteristics of topic-level temporal patterns for fake news detection.
- **Method:** We propose a framework for **F**orecasting **T**emporal **T**rends (**FTT**) to tackle temporal generalization issue in fake news detection.
- **Industrial Value:** We experimentally show that our FTT overall outperforms five compared methods while maintaining good compatibility with any neural network-based fake news detector.

2 Related Work

Fake News Detection. Fake news detection is generally formulated as a binary classification task between real and fake news items. Research on this task could be roughly grouped into content-only and social context-based methods. Content-only methods take the news content as the input including texts (Sheng et al., 2021), images (Qi et al., 2019), and videos (Bu et al., 2023), and aim at finding common patterns in news appearances. In this paper, we focus on textual contents but our method could be generalized to other modalities. Previous text-based studies focus on sentiment and emotion (Ajao et al., 2019; Ghanem et al., 2021), writing style (Przybyla, 2020), discourse structure (Karimi and Tang, 2019), etc. Recent studies address the domain shift issues across news channels and propose multi-domain (Nan et al., 2021; Zhu et al., 2022b) and cross-domain (Nan et al., 2022; Lin et al., 2022) detection methods. Zhu et al. (2022a) design a causal learning framework to remove the non-generalizable entity signals. Social context-based methods leverage crowd feedbacks (Kochkina et al., 2018; Shu et al., 2019; Zhang et al., 2021), propagation patterns (Zhou and Zafarani, 2019; Silva et al., 2021a), and social networks (Nguyen et al., 2020; Min et al., 2022), which have to wait for the accumulation of such social contexts.

Considering the in-time detection requirement,

[1] We denote the four quarters of a calendar year as Q1-Q4, respectively. For instance, Q1 stands for January through March.

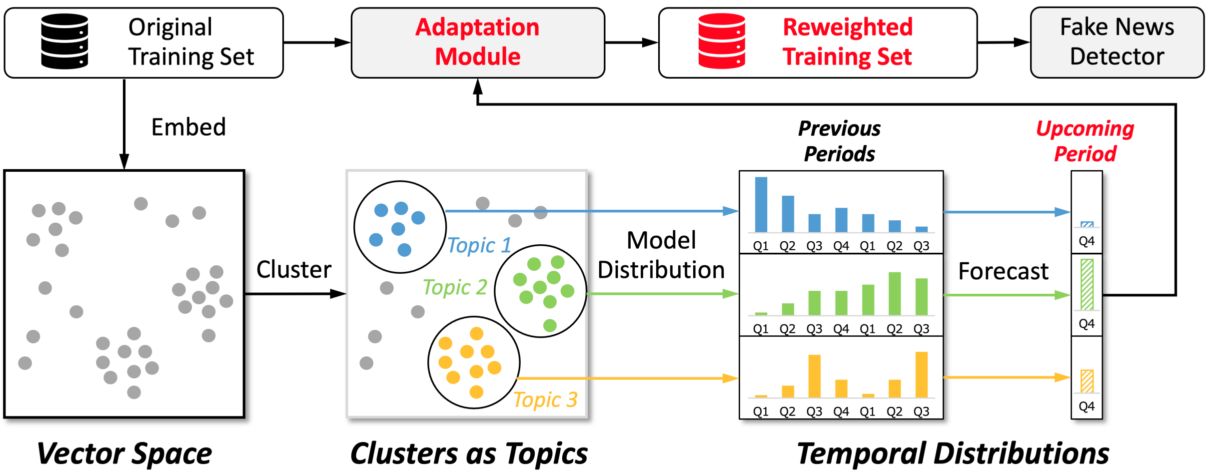

Figure 2: Architecture of the proposed **FTT** (**F**orecasting **T**emporal **T**rends) framework.

our proposed framework falls into the category of content-only methods, where we provide a new perspective for addressing the temporal generalization issue by forecasting temporal trends.

Temporal Generalization. The temporal generalization issue presents a situation in that models are trained on past data but required to perform well on unavailable and distribution-shifted future data. It has been addressed in a variety of applications such as review classification (Huang and Paul, 2019), named entity recognition (Rijhwani and Preotiuc-Pietro, 2020), and air quality prediction (Du et al., 2021). Recently, Gaspers et al. (2022) explore several time-aware heuristic-based instance reweighting methods based on recency and seasonality for an industrial speech language understanding scenario. Our work follows this line of instance reweighting, but we attempt to model the temporal patterns and forecast topic-wise distribution to better adapt to future data.

3 Proposed Framework

Our framework FTT is presented in Figure 2, where the instances from past consecutive time periods in the original training set are reweighted according to the forecasted topic-wise distribution for generalizing better in the upcoming time period. In the following, we first provide the problem formulation and subsequently, detail the procedures.

3.1 Problem Formulation

Given a dataset $\mathcal{D} = \{\mathcal{D}_q\}_{q=1}^{Q}$ consisting of Q subsets that contain news items from Q consecutive time periods, respectively, our goal is to train a model on $\{\mathcal{D}_q\}_{q=1}^{Q-1}$ that generalizes well on $\mathcal{D}_Q$. In $\mathcal{D}$, an instance is denoted as (x_i, y_i) where the ground-truth label $y_i = 1$ if the content x_i is fake.

In practice, we retrain and redeploy the fake news detector at a fixed time interval to reflect the effects of the latest labeled data. We set the interval as three months (i.e., a quarter) since a shorter interval does not allow sufficient accumulation of newly labeled fake news items. In the following, we set $\mathcal{D}_q$ as the subset corresponding to news in a quarter of a calendar year.

3.2 Step 1: News Representation

We first transform the news content into a vector space to obtain its representation, which will be used for similarity calculation in the subsequent clustering step. We employ Sentence-BERT (Reimers and Gurevych, 2019), which is widely used for sentence representation (e.g.,Shaar et al., 2020). For instance x_i, the representation vector is $\boldsymbol{x}_i \in \mathbb{R}^{768}$.

3.3 Step 2: Topic Discovery

We perform clustering on news items based on the representation obtained in Step 1 to group news items into distinct clusters which correspond to topics. Due to the lack of prior knowledge about the topic number, we adopt the single-pass incremental clustering algorithm which does not require a preset cluster number. We first empirically set a similarity threshold θ_{sim} to determine when to add a new cluster. When an item arrives, it is assigned to the existing cluster whose center is the nearest to it if the distance measured by cosine similarity is larger than θ_{sim}. Otherwise, it will be considered as an item on a new topic and thus be in a new independent cluster.

3.4 Step 3: Temporal Distribution Modeling and Forecasting

Based on the clustering results, we model the temporal distribution of different news topics and forecast the topic-wise distribution in the upcoming time period in this step. Note that we do not consider the clusters with news items less than the threshold θ_{count} since they are too small to present significant temporal patterns.

Modeling. Assuming that T topics are preserved, we first count the number of news items per quarter within each topic. The counts of the same quarter are then normalized across topics to obtain the quarterly frequency sequence of each topic (denoted as f). To model the temporal distribution, we adopt a decomposable time series model (Harvey and Peters, 1990) on the quarterly sequences and consider the following two trends (exemplified using Topic i):

1) General Trend. A topic may increase, decrease, or have a small fluctuation in terms of a general non-periodic trend (e.g., Topics 1 and 3 in Figure 1). To fit the data points, we use a piecewise linear function:

$$g_i(f_{i,q}) = k_i f_{i,q} + m_i, \tag{1}$$

where $k_i = k + \boldsymbol{a}(q)^T \boldsymbol{\delta}$ is the growth rate, $f_{i,q}$ is the frequency of Topic i in Quarter q, and $m_i = m + \boldsymbol{a}(q)^T \boldsymbol{\gamma}$ is the offset. k and m are initial parameters. $\boldsymbol{a}(q)$ records the changepoints of growth rates and offsets while $\boldsymbol{\delta}$ is the rate adjustment term and $\boldsymbol{\gamma}$ is a smoothing term.

2) Quarterly Trend. For topics having quarterly periodic trends like Topic 2 in Figure 1, we add four extra binary regressors corresponding to Q1~Q4 to inform the regression model the quarter that a data point in input sequence belongs to. For Topic i and Quarter q, we obtain the quarterly seasonality function $s_i(f_{i,q})$ by summing the four regression models.

Forecasting. We fit the model using the time series forecasting tool Prophet (Taylor and Letham, 2018) with the temporal distribution of topics from Quarter 1 to Quarter Q-1. To forecast the trend of Topic i in the upcoming Quarter Q, we sum up the two trend modeling functions:

$$p_i(f_{i,Q}) = g_i(f_{i,Q}) + s_i(f_{i,Q}). \tag{2}$$

3.5 Step 4: Forecast-Based Adaptation

Based on the topic-wise forecasts of frequency distribution in Quarter Q, we apply instance reweighting to the training set and expect the model trained using the reweighted set would better adapt to the future data in Quarter Q.

We first filter out topics that do not exhibit obvious regularity. Specifically, we remove the topics which have a mean absolute percentage error (MAPE) larger than a threshold θ_{mape} during the regression fitting process. For a Topic i in the preserved set Q', we calculate and then normalize the ratio between the forecasted frequency of Topic i $p_i(f_{i,Q})$ and the sum of all forecasted frequencies of the preserved topics:

$$w_{i,Q} = \mathrm{Bound}\left(\frac{p_i(f_{i,Q})}{\sum_{i\in Q'} p_i(f_{i,Q})}\right), \tag{3}$$

where Bound is a function to constrain the range of calculated weights. We set the weight smaller than θ_{lower} and larger than θ_{upper} as θ_{lower} and θ_{upper}, respectively, to avoid the instability during the training process. For those that are not included in Q', we set their weights as 1.

The new weight of the training set instances of Topic i, $w_{i,Q}$, corresponds to our forecasts of how frequent news items of this topic will emerge in the upcoming period Q. If the forecasted frequency of Topic i indicates a decreasing trend, the value will be smaller than 1 and thus instances of this topic will be down-weighted; conversely, if the forecasted distribution indicates an increasing trend, the value will be greater than 1 and the instances will be up-weighted. In the next step, we will show the reweighting process during training.

3.6 Step 5: Fake News Detector Training

Our framework FTT could be compatible with any neural network-based fake news detector. Here, we exemplify how FTT helps detectors' training using a pretrained BERT model (Devlin et al., 2019). Specifically, given an instance $\boldsymbol{x}_i$, we concatenate the special token [CLS] and $\boldsymbol{x}_i$, and feed them into BERT. The average output representation of non-padded tokens, denoted as $\boldsymbol{o}_i$, is then fed into a multi-layer perception (MLP) with a sigmoid activation function for final prediction:

$$\hat{y}_i = \mathrm{sigmoid}(\mathrm{MLP}(\boldsymbol{o}_i)). \tag{4}$$

Our difference lies in using the new weights based on the forecasted temporal distribution to increase or decrease the impact of instances during back-propagation. Unlike most cases that use an *average* cross-entropy loss, we minimize the *weighted* cross-entropy loss function during training:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} w_{i,Q}\mathrm{CrossEntropy}(y_i, \hat{y}_i), \quad (5)$$

where $w_{i,Q}$ is the new weight for instance x_i and y_i is its ground-truth label. N is the size of a mini-batch of the training set.

4 Evaluation

We conduct experiments to answer the following evaluation questions:

- **EQ1:** Can FTT bring improvement to the fake news detection model in temporal generalization scenarios?
- **EQ2:** How does FTT help with fake news detection models?

4.1 Dataset

Our data comes from a large-scale Chinese fake news detection system, covering the time period from January 2016 to December 2020. To meet the practical requirements, the data was divided by quarters based on the timestamp. Unlike the existing academic datasets (Shu et al., 2020; Sheng et al., 2022), the dataset is severely imbalanced. To avoid instability during training, we randomly undersampled the subset of each quarter to achieve a ratio of 1:1 between fake and real news. Identical to the real-world setting, we adopt a *rolling training* experimental setup. If we train a model to generalize well in the time period Q, the training, validation, and testing sets would be $\{\mathcal{D}_i\}_{i=1}^{Q-2}$, $\mathcal{D}_{Q-1}$, and $\mathcal{D}_Q$, respectively. If the target is $Q+1$, then the three subsets would be $\{\mathcal{D}_i\}_{i=1}^{Q-1}$, $\mathcal{D}_Q$, and $\mathcal{D}_{Q+1}$. Here we use the four quarterly datasets from 2020 as the testing sets and conduct experiments on the four sets separately.

4.2 Experimental Settings

Compared Methods. We compared our proposed FTT with five existing methods (including the vanilla baseline model), in which the second one is to remove non-generalizable bias and the last three are to introduce heuristic rules for adapting to future data.

- **Baseline** follows a normal training strategy where all training instances are equally weighted.
- **EANN$_T$** (Wang et al., 2018) is a model that enhances model generalization across events by introducing an auxiliary adversarial training task to prevent the model from learning event-related features. For fair comparison, we replaced the original TextCNN (Kim, 2014) with a trainable BERT as the textual feature extractor, and utilized publication year labels as the labels for the auxiliary task following Zhu et al., 2022a. We removed the image branch in EANN as here we focus on text-based fake news detection.
- **Same Period Reweighting** increases the weights of all training instances from the same quarter as the target data. It models the seasonality in the time series data.
- **Previous Period Reweighting** increases the weights of all training instances from the last quarter. It could capture the recency in the data distribution.
- **Combined Reweighting** combines the two reweighting methods mentioned above. The last three methods are derived from (Gaspers et al., 2022).

Implementation Details. We used a BERT model, `hfl/chinese-bert-wwm-ext` (Cui et al., 2021) implemented in HuggingFace's Transformer Package (Wolf et al., 2020) as the baseline fake news detection classifier. In the training process, we used the Adam optimizer (P. Kingma and Ba, 2015) with a learning rate of 2e-5 and adopted the early stop training strategy, and reported the testing performance of the best-performing model on the validation set. We employed grid search to find the optimal hyperparameters in each quarter for all methods. In Q1 and Q2, the optimal hyperparameters of FTT are $\theta_{sim} = 0.65$, $\theta_{count} = 30$, $\theta_{mape} = 0.8$, $\theta_{lower} = 0.3$, and $\theta_{upper} = 2.0$; and in Q3 and Q4, they are $\theta_{sim} = 0.5$, $\theta_{count} = 30$, $\theta_{mape} = 2.0$, $\theta_{lower} = 0.3$, and $\theta_{upper} = 2.0$.

We report the accuracy, macro F1 (macF1), and the F1 score for real and fake classes ($\mathrm{F1}_{\mathrm{real}}$ and $\mathrm{F1}_{\mathrm{fake}}$).

4.3 Performance Comparison (EQ1)

Table 1 shows the overall and quarterly performance of the proposed framework and other methods. We observe that:

2020	Metric	Baseline	EANN_T	Same Period Reweighting	Prev. Period Reweighting	Combined Reweighting	FTT (Ours)
Q1	macF1	0.8344	0.8334	0.8297	0.8355	0.8312	**0.8402**
	Accuracy	0.8348	0.8348	0.8301	0.8359	0.8315	**0.8409**
	F1_{fake}	0.8262	0.8181	0.8218	0.8274	0.8237	**0.8295**
	F1_{real}	0.8425	0.8487	0.8377	0.8435	0.8387	**0.8509**
Q2	macF1	0.8940	0.8932	0.8900	0.9004	0.8964	**0.9013**
	Accuracy	0.8942	0.8934	0.8902	0.9006	0.8966	**0.9014**
	F1_{fake}	0.8894	0.8887	0.8852	0.8953	0.8915	**0.8981**
	F1_{real}	0.8986	0.8978	0.8949	**0.9055**	0.9013	0.9046
Q3	macF1	0.8771	0.8699	0.8753	0.8734	0.8697	**0.8821**
	Accuracy	0.8776	0.8707	0.8759	0.8741	0.8707	**0.8827**
	F1_{fake}	0.8696	0.8593	0.8670	0.8640	0.8582	**0.8743**
	F1_{real}	0.8846	0.8805	0.8836	0.8829	0.8812	**0.8900**
Q4	macF1	0.8464	0.8646	0.8464	0.8429	0.8412	**0.8780**
	Accuracy	0.8476	0.8647	0.8476	0.8442	0.8425	**0.8784**
	F1_{fake}	0.8330	0.8602	0.8330	0.8286	0.8271	**0.8707**
	F1_{real}	0.8598	0.8690	0.8598	0.8571	0.8553	**0.8853**
Average	macF1	0.8630	0.8653	0.8604	0.8631	0.8596	**0.8754**
	Accuracy	0.8636	0.8659	0.8610	0.8637	0.8603	**0.8759**
	F1_{fake}	0.8546	0.8566	0.8518	0.8538	0.8501	**0.8682**
	F1_{real}	0.8714	0.8740	0.8690	0.8723	0.8691	**0.8827**

Table 1: Performance of the baseline method, four existing methods, and our method in fake news detection. The best result in each line is **bolded**.

1) FTT outperforms the baseline and four other methods across all quarters in terms of most of the metrics (the only exception is F1_{real} in Q2). These results demonstrate its effectiveness.

2) The average improvement of F1_{fake} is larger than that of F1_{real}, suggesting that our method helps more in capturing the uniqueness of fake news. We attribute this to the differences in temporal distribution fluctuation: fake news often focuses on specific topics, while real news generally covers more diverse ones. This makes the topic distribution of fake news more stable, which allows for better modeling of topic-wise distributions.

3) The three compared reweighting methods show inconsistent performances. In some situations, the performance is even lower than the baseline (e.g., Same Period Reweighting in Q1). We speculate that the failure is caused by the complexity of the news data. Considering the rapidly-evolving nature of news, single heuristic methods like recency and seasonality could not fast adapt to future news distribution. In contrast, our FTT performs topic-wise temporal distribution modeling and next-period forecasting and thus has a better adaption ability.

Subset of the test set	Metric	Baseline	FTT (Ours)
Existing Topics	macF1	0.8425	**0.8658**
	Accuracy	0.8589	**0.8805**
	F1_{fake}	0.7997	**0.8293**
	F1_{real}	0.8854	**0.9023**
New Topics	macF1	0.8728	**0.8846**
	Accuracy	0.8729	**0.8846**
	F1_{fake}	0.8730	**0.8849**
	F1_{real}	0.8727	**0.8843**

Table 2: Breakdown of the performance on the testing set according to the existence of their topics.

4.4 Result Analysis (EQ2)

Statistical Analysis. To analyze how FTT improves fake news detection performance, we analyze the testing instances by recognizing their topics. Specifically, we run the single-pass incremental clustering algorithm used in Step 2 again on the testing instances based on the clusters on the training set. If a news item in the testing set could be clustered into an existing cluster, it will be recognized as an item of the existing topics; otherwise, it will be in a new topic. Based on the results, we show the breakdown of the performance on the testing set in Table 2. Compared with the baseline, our framework achieves performance improvements on both the Existing Topics and the New Topics subsets. This could be attributed to our reweighting

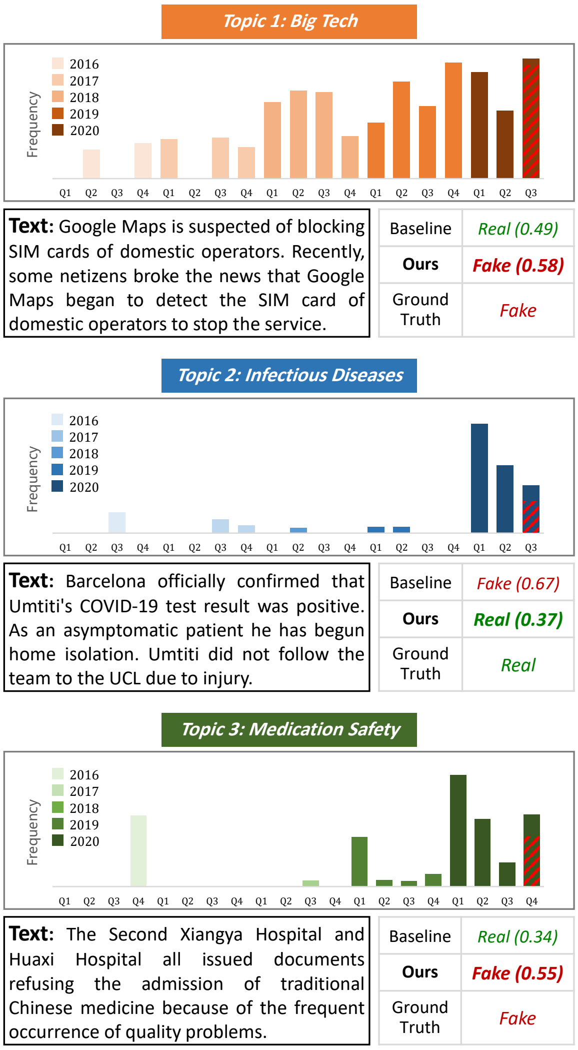

Figure 3: Three cases from the testing set. The forecasts by FTT about the frequency of the topics in the upcoming quarter are highlighted with red dashed bars. The case texts are translated from Chinese into English.

strategy where we not only increase the weights of news items belonging to a topic of an increasing trend but also decrease the weights of those belonging to the fading topics. With such a design, the model will be more familiar with news items in existing topics and more generalizable to news items in new topics.

Case Study. Figure 3 shows three cases from the testing set. According to the forecasted results of the frequencies of these topics in the testing time period, our framework assigns positive weights (greater than 1) to items in these topics. After training on the reweighted set, the detector flips its previously incorrect predictions. In Topic 1, the frequency of Big Tech-related news items demonstrated an increasing trend over time. FTT captures this pattern and provides a forecast close to the true value for the target quarter. In Topic 2, there is an explosive growth of Infectious Diseases-related news items in early 2020, followed by sustained high frequency in the subsequent quarters. FTT successfully captures this change. In contrast to the other two topics, the frequency of Medication Safety-related news items in Topic 3 exhibits both an overall increasing trend and a certain periodic pattern since 2019, which roughly follows a “smiling curve” from Q1 to Q4 in a single year. FTT effectively models both of these patterns and helps identify the importance of news items in this topic for the testing time period.

5 Conclusion and Future Work

We studied temporal generalization in fake news detection where a model is trained with previous news data but required to generalize well on the upcoming news data. Based on the assumption that the appearance of news events on the same topic presents diverse temporal patterns, we designed a framework named FTT to capture such patterns and forecast the temporal trends at the topic level. The forecasts guided instance reweighting to improve the model’s generalizability. Experiments demonstrate the superiority of our framework. In the future, we plan to mine more diverse temporal patterns to further improve fake news detection in real-world temporal scenarios.

Limitations

We identify the following limitations in our work:

First, our FTT framework captures and models topic-level temporal patterns for forecasting temporal trends. Though the forecasts bring better temporal generalizability, FTT could hardly forecast the emergence of events in new topics.

Second, FTT considers temporal patterns based on the topic-wise frequency sequences to identify patterns such as decrease, periodicity, and approximate stationery. There might be diverse patterns that could not be reflected by frequency sequences.

Third, limited by the scarcity of the dataset that satisfies our evaluation requirements (consecutive time periods with a consistent data collection criterion), we only performed the experiments on a Chinese text-only dataset. Our method should be further examined on datasets of other languages and multi-modal ones.

Acknowledgements

The authors thank anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (62203425), the Zhejiang Provincial Key Research and Development Program of China (2021C01164), the Project of Chinese Academy of Sciences (E141020), and the Innovation Funding from Institute of Computing Technology, Chinese Academy of Sciences (E161020).

References

Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2507–2511. IEEE.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. *arXiv preprint arXiv:2302.03242*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

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. ACL.

Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. 2021. AdaRNN: Adaptive learning and forecasting for time series. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 402–411. ACM.

Judith Gaspers, Anoop Kumar, Greg Ver Steeg, and Aram Galstyan. 2022. Temporal generalization for spoken language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 37–44. ACL.

Bilal Ghanem, Simone Paolo Ponzetto, Paolo Rosso, and Francisco Rangel. 2021. FakeFlow: Fake news detection by modeling the flow of affective information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 679–689. ACL.

Andrew C Harvey and Simon Peters. 1990. Estimation procedures for structural time series models. *Journal of forecasting*, 9(2):89–108.

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. ACL.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. 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 3432–3442. ACL.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. ACL.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413. ACL.

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. ACL.

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. ACM.

Ahmadreza Mosallanezhad, Mansooreh Karami, Kai Shu, Michelle V. Mancenido, and Huan Liu. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM Web Conference 2022*, page 3632–3640. ACM.

Yida Mu, Kalina Bontcheva, and Nikolaos Aletras. 2023. It's about time: Rethinking evaluation on rumor detection benchmarks using chronological splits. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 736–743. ACL.

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 and Knowledge Management*. ACM.

Qiong Nan, Danding Wang, Yongchun Zhu, Qiang Sheng, Yuhui Shi, Juan Cao, and Jintao Li. 2022. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings* *of the 29th International Conference on Computational Linguistics*, pages 2834–2848. ICCL.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 1165–1174. ACM.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Piotr Przybyla. 2020. Capturing the style of fake news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 490–497. AAAI Press.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining*, pages 518–527. IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. ACL.

Shruti Rijhwani and Daniel Preotiuc-Pietro. 2020. Temporally-informed analysis of named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7605–7617. ACL.

Yoel Roth. 2022. The vast majority of content we take action on for misinformation is identified proactively. `https://twitter.com/yoyoel/status/1483094057471524867`.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618. ACL.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556. ACL.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern- and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 1640–1650. ACM.

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. ACM.

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

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.

Amila Silva, Yi Han, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021a. Propagation2Vec: Embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management*, 58(5):102618.

Amila Silva, Ling Luo, Shanika Karunasekera, and Christopher Leckie. 2021b. Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 557–565. AAAI Press.

Sean J Taylor and Benjamin Letham. 2018. Forecasting at scale. *The American Statistician*, 72(1):37–45.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 849–857. ACM.

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. ACL.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476. ACM.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD Explorations Newsletter*, 21(2):48–60.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022a. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125. ACM.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2022b. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14.