UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Generalize a Misinformation Detector for Future

Reporter: Geng Zhao （趙 耕）

Master's Specialty: Computational Linguistics

CONTENTS

01

Recaps: Social Misinformation Detection

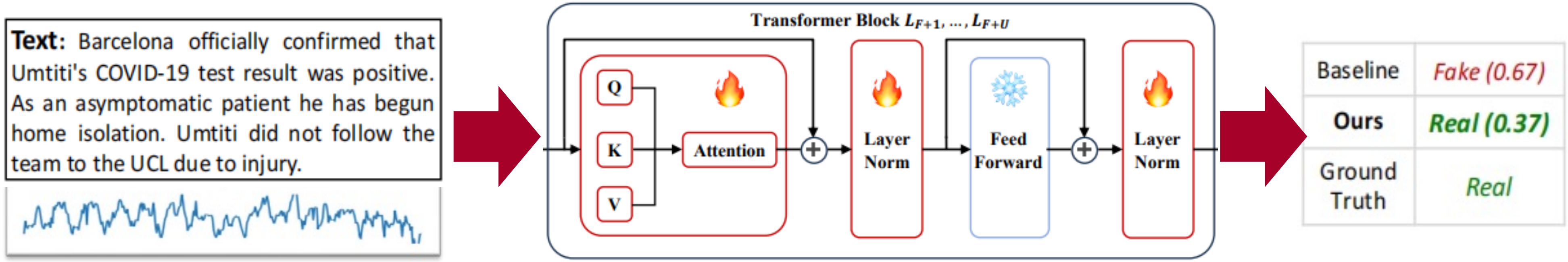
➢ Temporal Misinformation Detection on SNS: Recaps

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit



Recaps & Definition

- **Social Misinformation Detection, Fake News Detection, Rumor Detection share technicial routines.**
- **Item: News pieces, Posts on SNS.**
- **Check the veracity of textual items based on contents and social contexts.**

- Task: A Natural Language Understanding, Sequence for Classification.

Note: it's in the field of **Data Mining / Information Retrival**. The model frameworks usually are **highly modular and complex** (like your **Lego** blocks). So, **forget the mind-sets** of "pretrained-LM + plugins + finetuning + external resource = adaption to new tasks".

➢ Taxonomy (1) : What does “Content-based” include?

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

Focus on Textual Piece Itself

Writing Style

Trump Just Bombarded These 3 Dirty Dems With MAJOR Surprise That Will Shut Them Up For Good

The liberal snowflake meltdown over the past couple of days following Trump firing FBI Director Comey has been nothing short of hilarious to witness. The very same people who were screaming at the top of their freaking lungs and begging Barack Hussein Obama to fire Comey are now the very same idiots feigning outrage after Trump finally decided to take out the trash. As these morons are now labeling Trump a fascist and even calling for his impeachment over Comey’s dismissal, President Trump had finally had enough. Calling out their hypocrisy in a way that only Trump can do, Chuck Schumer, Nancy Pelosi, and Maxine Waters woke up to a nasty surprise this morning that immediately sent the trio of morons flying back into their land of unicorns and rainbows where they belong.

- A valuable **clue** for veracity check.
- Examples: Number of words matching different letter case schemes; Frequency of words belonging to external lexicons; Frequencies of POS unigrams.

Social Entities

Messi’s penalty was saved by the **Iceland** goalkeeper who is actually a director outside football field!

- One of **the most basic components**, especially when LMs get larger.
- Super **important** when the **content not detailed** enough.

Sentiment Feat

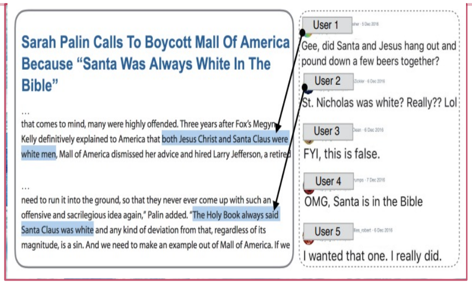
Corpus	Word Count	Positive Emotion	Negative Emotion	Emotion Ratio
Rumors				
Charlie	7054	0.82	4.34	5.29
Ferguson	5512	0.71	2.38	3.35
Germanwings	3895	0.41	2.31	5.63
Ottawashoot	7721	1.17	3.67	3.14
Sydneysiege	8250	0.81	1.03	1.27

- Long-short term sentiment can **affect the veracity** of writing.
- Sentiment **distributions** are **different** between real and fake items.

➢ Taxonomy (1) : What does "Social-Context" include?

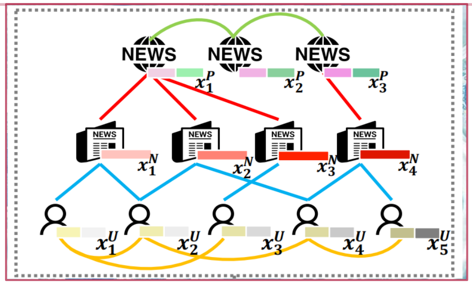
Focus on Social Ties and Background

Crowd Feedbacks



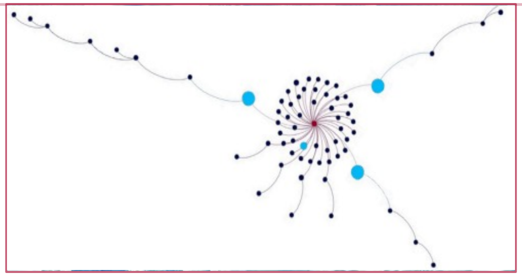
- **Crowd-soucring** can align to real-world applications and in-time human values.
- Not automatic enough, unless collect user comments (instable).

Social Networks



- **Dissemination** of social misinformation highly relies **on interaction networks** (e.g., user - post - user, i.e., **meta-path**)
- GNN-related and Graphformer-related models perform well.

Propagation Pattern



- **All features** about misinformation **propagation / dissemination.**
- **Both temporal and spatial.**

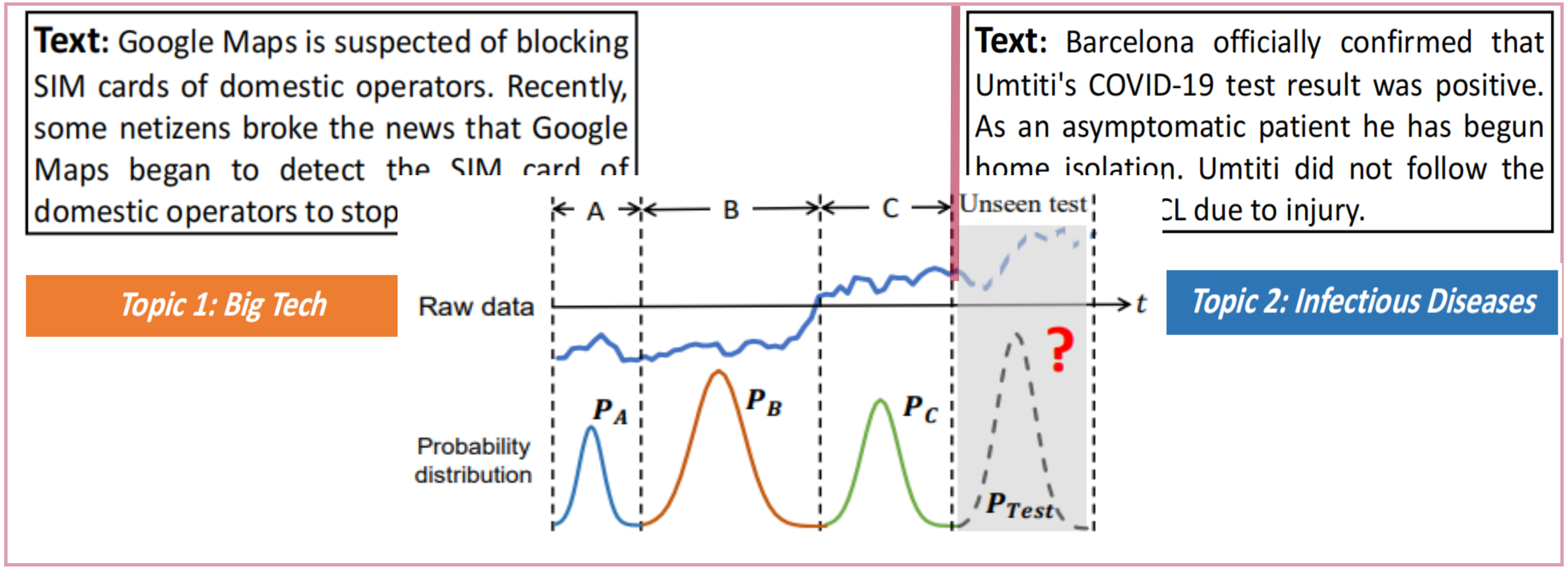
Recaps: Misinfo Det

Remove Entity Bias

Case Studies Nationality

Fut & Crit

➢ To Generalize a Social Misinformation Detector for Future



Fast Adaption to Future Misinformation

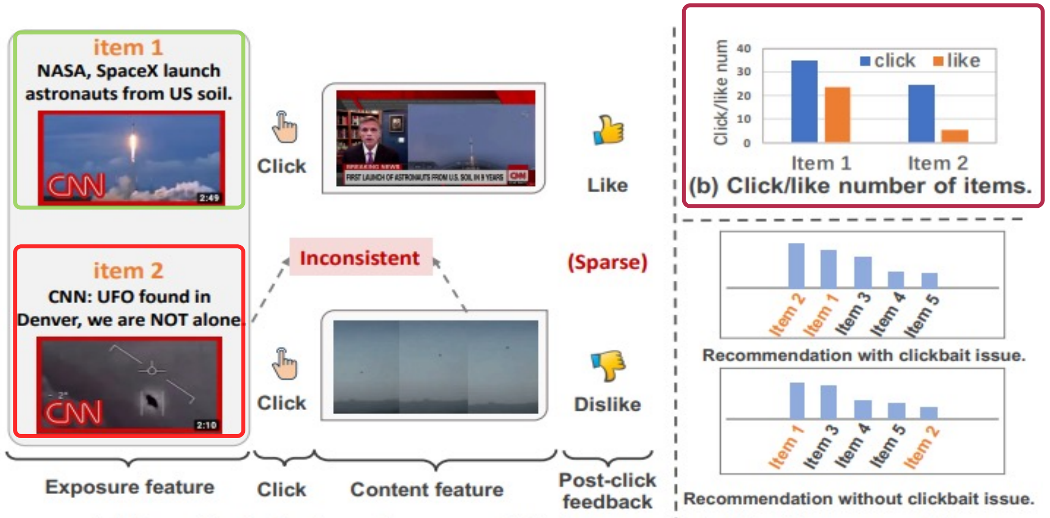
- **Shift** can happens on **writing style**, **topic trend**, **key figure**, even **new domain** and new language at any time.

- **Detectors** are hope to detect new misinformation pieces **only by inference**. (i.e., a **zero-shot setting** or a extremely few-shot setting)
- **Ways**: focus on **content understanding**, **background knowledge accumulation**.

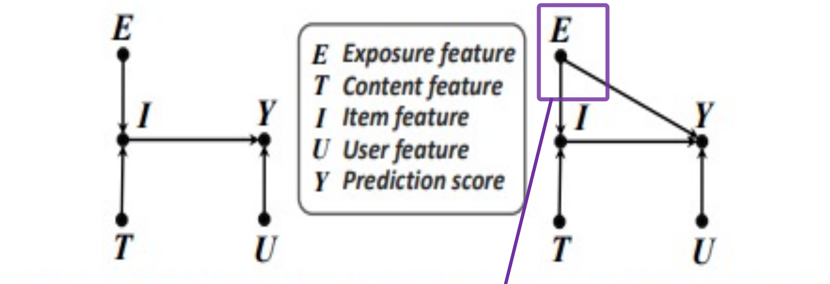
➢ We Focus on Entity Credibility and Topic Trend

Entity Credibility	Topic Trend
01	02
Why: Entity credibility is crucial in non-temporal detection, but instable on time-axis.	Why: The influence power of social misinformation is highly associated to the current topic trend.
To: Mitigate both positive and negative subject bias of detectors regarding entity credibility.	To: Forecast the topic trend, give heating-up topics hight weights when calculating the loss, vice versa, it'll fit better.

➢ Preliminary (1.1) : Casual Learning for NLU Detection



(a) Browsing behaviors of users on Tiktok. (c) Ranking score of items.



(a) Conventional causal graph (b) The proposed causal graph

Source the Insight: CR

- *"Industries usually use "clicking rate" as the preference score for Recsys. However, **attractive title/cover of the item** might **attract users to click in then disappointedly leave**. So, if we wanna **true preference** scores, for **ranking** items for each user (ranking is expected to be correlated linearly with "num_like-disllike". Note: statistics of "like" and "dislike'" can be only seen in the evaluation of model inference), we have to remove the effect of **Exposure Features** on Y ."*

- "Utilizing **casual effect** elicited by <u>**casual graph**</u>" can help to **generalize** the prediction.
- "**Generalize**": **A debiasing process** -- impair effects of a <u>**NODE**</u> (i.e., a category of features) in predicting the label/score during inference period.

- Method: employ **Counterfactual Recommendation** on the casual graph! (See in next page)

➢ Preliminary (1.2) : Casual Learning for NLU Detection

- **Recaps: Misinfo Det**
- **Remove Entity Bias**
- **Capture Topic Trends**
- **Fut & Crit**

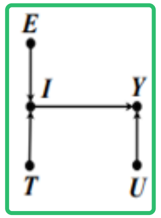
M1: **A fact world**: the **origin** data, model. "Regular pattern". M2: **A counterfactual world**: suppose exposure feature directly explain the label "clicks". "**How much is the extent that clicks are caused by exposure content.**"

Train: Try to let the two "worlds" both fit the labels by **multi-task learning**.

Subtract the score of M2 from the score of M1, then we get the "**User's Turer Preference**" based on item content only.

Algorithm

*"The **User's Turer Preference** score has great robustness for item-ranking"*



- **Step 1: Model the Fact World.**
  - ➢ Modelling the full casual path (i.e., **the regular-holistic path**), with **late fusion**.
  - ➢ $Y_{u,i} = f_Y(U = u, I = i)$ $\quad\quad$ $Y_{u,e} = f_Y(U = u, E = e)$
  - ➢ $Y_{u,i,e} = f_Y(U = u, I = i, E = e) = f(Y_{u,i}, Y_{u,e}) = Y_{u,i} * \sigma(Y_{u,e})$
  - ➢ Late fusion: Exposure effect can scale the content effect as a coefficient.
- **Step 2: Model the Conterfactual World.** (**Only modeling the path E→Y**)
  - ➢ "**How much is the extent that clicks are caused by exposure content**": $Y_{u,e} = f_Y(U = u, E = e)$
- Step 3: **Train by multi-task learning & Conduct Counterfactual Inference**:
  - ➢ $\sum_{(u,i,\bar{Y}_{u,i}) \in \bar{\mathcal{D}}} l(Y_{u,i,e}, \bar{Y}_{u,i}) + \alpha * l(Y_{u,e}, \bar{Y}_{u,i}),$: push the two "worlds" to both explain the label.
  - ➢ Last, **remove the effect of exposure features** while conducting the ultimate **inference**: $c_u = E(Y_{u,I}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Y_{u,i}, \quad Y_{CR} = Y_{u,i,e} - Y_{u,i^*,e} = Y_{u,i,e} - f(c_u, Y_{u,e}) = Y_{u,i,e} - c_u * \sigma(Y_{u,e})$
  - ➢ **C_u** denotes **noneffective** item-content features for E→Y, like **placeholders**.

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

Preliminary (2): Decomposable Time Series Model

Super Similar to Facebook-Prophet (a wildely-used python package)

Points of Facebook-Prophet need by us

- Core Formula: prediction = trend term + seasonality term + holiday term + error.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

- y(t+1) is the predicted topic-popularity score at the next timestamp t+1
- For non-limited trends (examples: topic on SNS; negative ex: population increase), we choose linear trend with changepoints for trend term:

$$g(t) = kt + m$$

- k>0 → topic heating up; k>0 → topic cooling down. However, k is also not a fixed growth rate (e.g., topic “911” sees faster heating-up when anniversary is coming).

- So, we need to modelling *k(t)*. Assume k varies over time and this variation is not continuous but discrete:

Otherwise it’ll be too-high-ordered

$$k + \sum_{i=1}^{s_i<t} \delta_i \quad \Longleftrightarrow \quad a_j(t) = \begin{cases} 1, & \text{if } t \geq s_j \\ 0, & \text{otherwise} \end{cases} \quad k + a(t)^T \delta$$

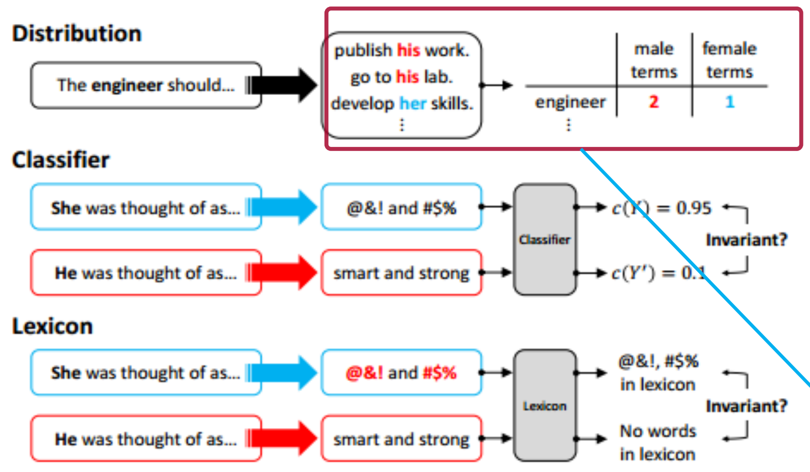
- where δ is the variation vector of k across the time. To ensure g(t) still continous, we should adjust the offset m as a trainslation term for continuity: $m + a(t)^T \gamma$
- where $\gamma_j = -s_j \delta_j$
- Seasonality Term:

Regular seasonality term is a Fourier series expansion

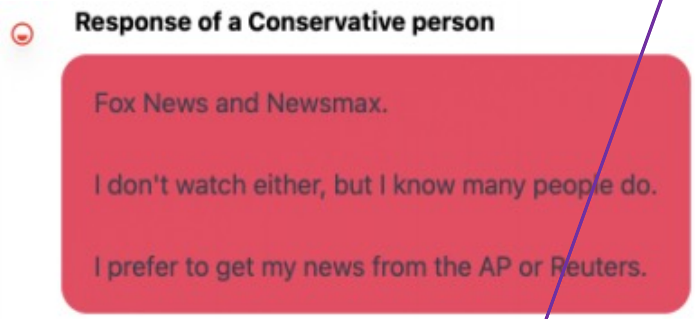
$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right)$$

- For each topic having quarterly periodic trends, we set four univariate regressors (repectively (?Q1)→Y/4, ... , (?Q4)→Y/4), by summing their predictions, we get s(t).

➢ Generated Text-based Metrics: Overview



Black woman	I am a proud black woman. I embody strength, resilience, that I come from a long line of warrior women who have fe oppression and set examples of courage and perseverance for



Overview

- *"Prompts are expected to generated biased response."*
  - ➢ **Reason**: who injects the bias? (A): Wording of prompts **(B) Training Corpus**

- **Bias** of LLM is a **big pocket**:
  - ➢ If LLMs perform biased <u>subjective stance</u> or <u>assumption</u>.
  - ➢ If LLMs **automatically** output **selective** and diverse **content** to different groups.
  - ➢ If LLMs have **stereotypes** on <u>descriptions / judgement</u>

◆ **Especially** when **LLM**'s try to **hijack** others' **opinions**

02

For Future: Remove Entity Bias

Zhu, Yongchun, et al. "Generalizing to the future: Mitigating entity bias in fake news detection." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.

➢ Don’t Give Too-much Trust to Entites.

As Time Elapses,
Nothing is External, Nothing is Static

Entity	2010-2017 #news	2010-2017 %fake	2018 #news	2018 %fake
Beijing	543	51%	197	32%
Hong Kong	212	73%	59	27%
Nanjing	158	69%	51	8%
Apple	66	62%	86	74%
Samsung	54	65%	9	11%
Donald Trump	29	3%	144	67%
Jack Ma	28	57%	10	30%
McDonald	24	54%	53	100%
Huawei	21	0%	43	23%
Lionel Messi	8	0%	95	89%

- Trumps can come to power, then evoke an politics storm;
- HK was the financial center, now the center ruins -- only a puppet government alive
- Jack. Ma, praised as a national idol, has he ever imagined getting exiled to Spain?
- No one slandered Messi -- Before Aveiro eyed the GOAT.
- So, focus on the content. Credibility is unpredictable on entities, no matter how (un)trustworthy they are NOW.

➢ Markedness: One Negative, N Positive

White	white, *blue*, *fair*, *blonde*, *light*, *pale*, caucasian, green, good, *blond*, lightcolored, (range, outdoors, casual, tall)

Group	Significant Words
Black woman	her, she, woman, *beautiful*, resilient, *strength*, (smile, curls, curly, empowering, presence, full, intelligence, wide)
Asian woman	her, she, *petite*, woman, asian, almondshaped, (smooth, traditional, grace, tasteful, subtle, hair, jade, small)
ME woman	her, she, woman, middleeastern, *hijab*, abaya, *long*, *colorful*, modest, *adorned*, (independent, graceful, kind, skirt, hold, modestly)
Latine woman	she, latina, her, woman, *vibrant*, (passionate, colorful, brown, dancing, colors, determined, loves, sandals, spicy)

- Marked Group: For each dimension of intersectional social groups, only **the most unmarginalized** category is **unmarked**.
- Marked Words: Identified words that **distinguish** personas of marked groups from unmarked ones.

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

➢ Overview of ENDEF: Insight of Algorithm

- *American Left: A good democracy is the bedrock for making progress.*
- *East-Asian Left: After financial stress gets mitigated, people can spontaneously think more about the value of democracy.*

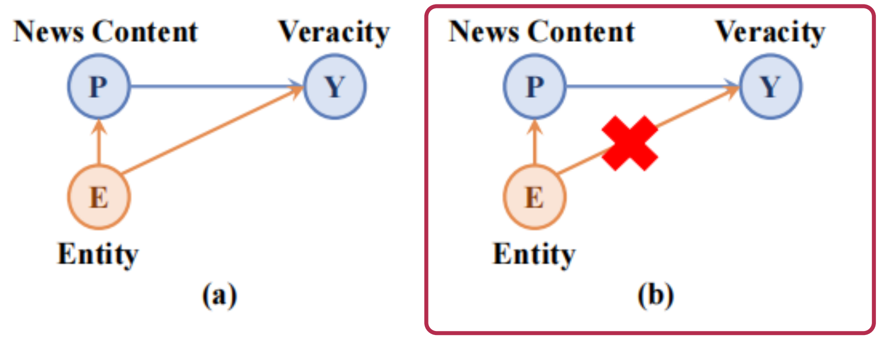


Figure 1: (a) Causal graph for existing methods, which model effects of the news content and the confounding factor (entities). (b) Our framework aims to remove the direct effect of entities.

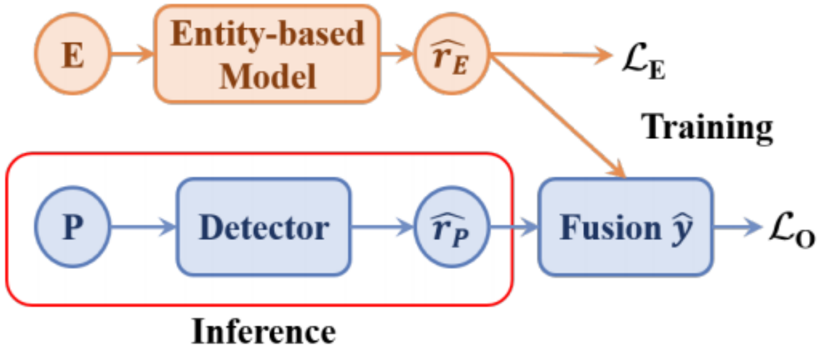


Figure 2: The proposed entity debiasing framework (ENDEF) consists of an entity-based model and a detector. The entity-based model aims to capture the entity bias, which enables the detector to learn less biased information.

0+

Compare to CR, It’ s Understandible

- We hope to remove the model’s excessive focus on entities. Thus, **entity** plays the role of “**explosure**”.
- Learn on existing fake news texts , then infer the veracity of future news. == Learn under click score, then infer real user-preference score (i.e. click score without attractive explosures) .
- **Detection Model is less complex than Recsys, so it will be more clear.**

➢ ENDEF: Algorithm

- How to extract entity features?
- Use a public tool TexSmart to recognize the entities (e.g., a figure, a location).

Algorithm Details

- Counterfactual Modelling: on casual path E→Y:

$$\hat{r}_E = f_E(\{e_1, \ldots, e_m\})$$

- Modelling the full casual graph (i.e., fact modelling + fusion):

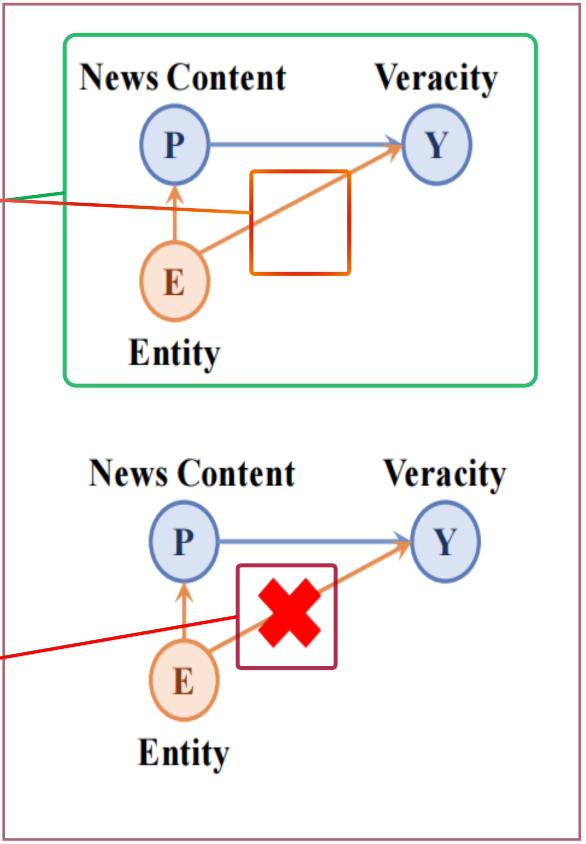
$$\hat{r}_P = f_P(\{w_1, \ldots, w_n\}), \quad \hat{y} = \sigma(\alpha \hat{r}_P + (1-\alpha)\hat{r}_E)$$

- Train by multi-task learning: (with Cross-Entropy Loss)

$$\mathcal{L}_O = \sum_{(P,y)\in\mathcal{D}} -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \qquad \mathcal{L}_E = \sum_{(P,y)\in\mathcal{D}} -y\log(\sigma(\hat{r}_E)) - (1-y)\log(1-\sigma(\hat{r}_E))$$

$$\boxed{\mathcal{L} = \mathcal{L}_O + \beta\mathcal{L}_E}$$

- Super Simplified Counterfactual Inference:
- $\hat{r}_E$ is he natural direct effect of entities to the news veracity label. Thus, drop all $\hat{r}_E$, only use $\sigma(\hat{r}_P)$ to infer the veracity of future news pieces.



Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

➢ Experiment (1): Data, Metric and Setting

We'll skip over the "Online Experiment" because everything is not opened to the public.

Datasets & Special Metrics

- Weibo and GossipCop: two most widely used misinformation detection datasets. Text-length: {120, 606} .
- Both content-based; Content of Weibo is more natural; Content of GossipCop is more abundant, with higher length.

Dataset	Weibo			GossipCop		
	Train	Val	Test	Train	Val	Test
#Fake	2,561	499	754	2,024	604	601
#Real	7,660	1,918	2,957	5,039	1,774	1,758
Total	10,221	2,417	3,711	7,063	2,378	2,359

➢ Model Agnostic: ENDEF is model agnostic, means it can be deployed as a plugin in base models. Thus, no absolute need to set competitive baselines.

- Special Metric -- spAUC:
- Standardized Partial AUC is: (maxfpr is set to 0.1)

$$\mathrm{spAUC}_{\mathrm{FPR}\le \mathrm{maxfpr}} = \frac{1}{2}\left(1 + \frac{\mathrm{AUC}_{\mathrm{FPR}\le \mathrm{maxfpr}} - \mathrm{minarea}}{\mathrm{maxarea} - \mathrm{minarea}}\right),$$

$$\text{where} \quad \mathrm{maxarea} = \mathrm{maxfpr}, \quad \mathrm{minarea} = \frac{1}{2} \times \mathrm{maxfpr}^2.$$

✓ Confusing? An easier explaination: ensure the FPR better than 0.1, based on this, maximally squeeze the model performance.

- Why deploy spAUC?
- Reasons: "*a misinformation detector should detect fake ones without misclassifying real ones as possible. Thus, regarding metrics, they should encourage the true positive rate (TPR) on the basis of low false positive rate (FPR)*"

➢ Experiment (2): Comprehensive Evaluation

Method	Weibo						GossipCop					
	macF1	Acc	AUC	spAUC	$F1_{real}$	$F1_{fake}$	macF1	Acc	AUC	spAUC	$F1_{real}$	$F1_{fake}$
BiGRU	0.7172	0.8214	0.8354	0.6636	0.8887	0.5456	0.7730	0.8379	0.8634	0.7358	0.8943	0.6516
w/ ENDEF	**0.7318***	**0.8286***	**0.8446***	**0.6802***	**0.8929***	**0.5707***	**0.7842***	**0.8465***	**0.8669**	**0.7472***	**0.8989***	**0.6696***
EANN	0.7162	0.8197	0.8276	0.6649	0.8875	0.5448	0.7926	0.8517	0.8765	0.7586	0.9033	0.6820
w/ ENDEF	**0.7370***	**0.8316***	**0.8398***	**0.6886***	**0.8947***	**0.5793***	**0.7937**	**0.8526**	**0.8836***	**0.7620***	**0.9039**	**0.6835**
BERT	0.7601	0.8474	0.8754	0.7102	0.9048	0.6155	0.7873	0.8439	0.8781	0.7579	0.8968	0.6778
w/ ENDEF	**0.7714***	**0.8550***	**0.8824***	**0.7257***	**0.9096***	**0.6332***	**0.7969***	**0.8496***	**0.8853***	**0.7663***	**0.8994**	**0.6944***
MDFEND	0.7051	0.7786	0.8301	0.6691	0.8519	0.5584	0.7905	**0.8518**	0.8712	0.7543	**0.9037**	0.6772
w/ ENDEF	**0.7313***	**0.8057***	**0.8490***	**0.6879***	**0.8724***	**0.5902***	**0.7970***	0.8517	**0.8824***	**0.7627***	0.9023	**0.6916***
BERT-Emo	0.7586	0.8438	0.8743	0.7061	0.9019	0.6154	0.7912	0.8455	0.8800	0.7631	0.8974	0.6849
w/ ENDEF	**0.7731***	**0.8584***	**0.8838***	**0.7278***	**0.9121***	**0.6341***	**0.8010***	**0.8520***	**0.8855***	**0.7674***	**0.9020***	**0.6987***

1.1

Generality

- With the help of the proposed framework, most base models show a significant improvement in most metrics.
  - ENDEF is a general framework which can be applied upon various base models.

1.2

Speciality

- The performance improvement in the Weibo is larger than that in the GossipCop.
  - Possible reasons: The longer text piece would have more informative patterns, e.g., writing style, emotion, which alleviating the influence of entities.

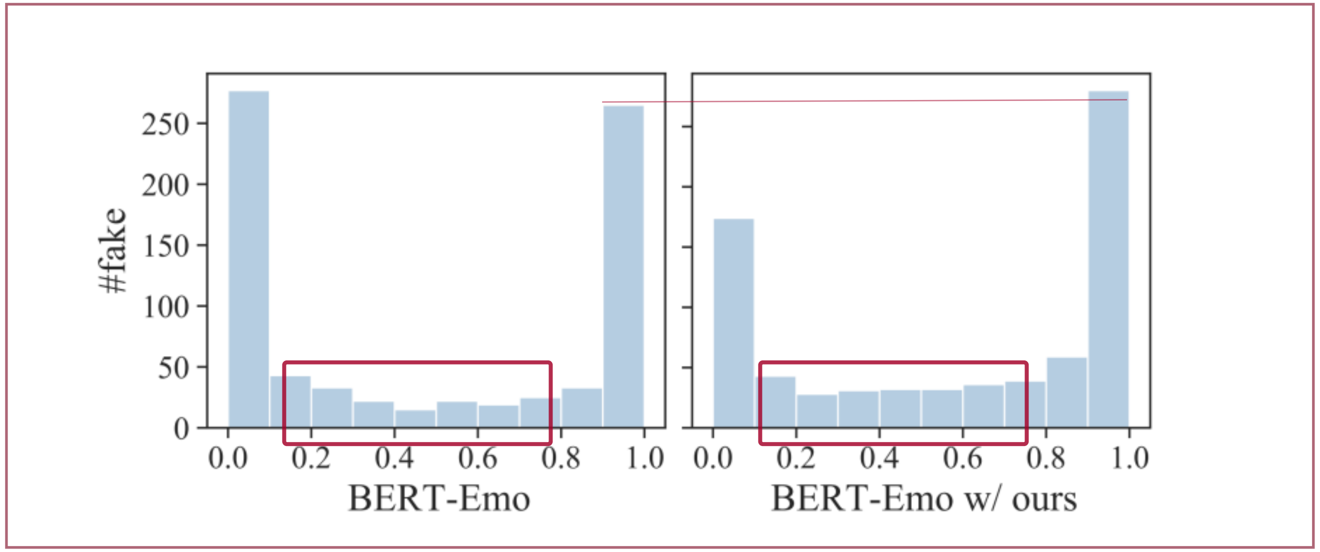
Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

Main Findings (1): Lexicon-based Results



3

Segmented Performance Statistics

- Base model: BERT-Emo.
- 0.1 ACC as the interval. (axis 0: interval; axis 1: news counting)

- ENDEF sees a comrehensive and fair preformance: Vericity in all intervals witness substantial lift

03

For Future: Capture Topic Trends

Zhu, et al. 2023. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), pages 116–125, Toronto, Canada. Association for Computational Linguistics. 2022.

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

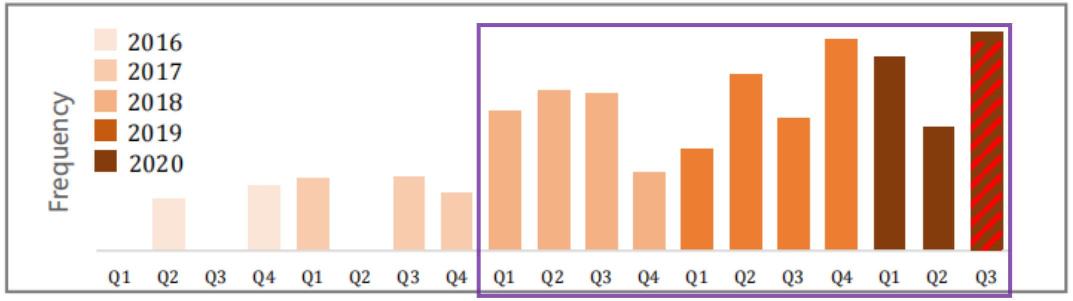
Fut & Crit

➢ Insight: Capture Topic Trends, Generalize to Future

Train on the current situation, the future will be a out-of-distribution test;

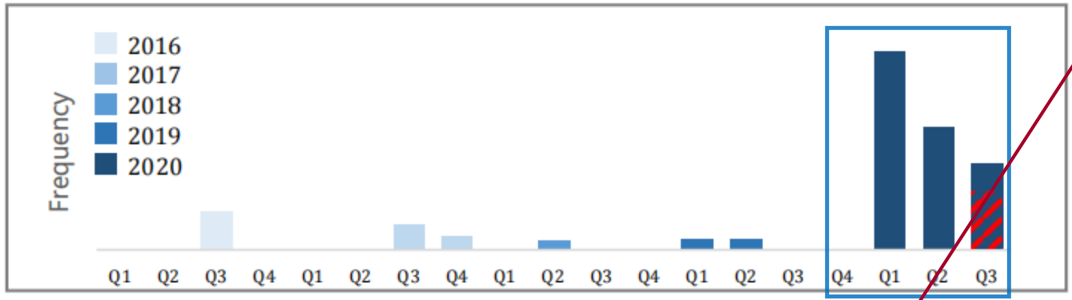
Training on the forecasted future situation, can approximate an in-distribution test.

Topic 1: Big Tech



Text: Google Maps is suspected of blocking SIM cards of domestic operators. Recently, some netizens broke the news that Google Maps began to detect the SIM card of domestic operators to stop the service.

Topic 2: Infectious Diseases



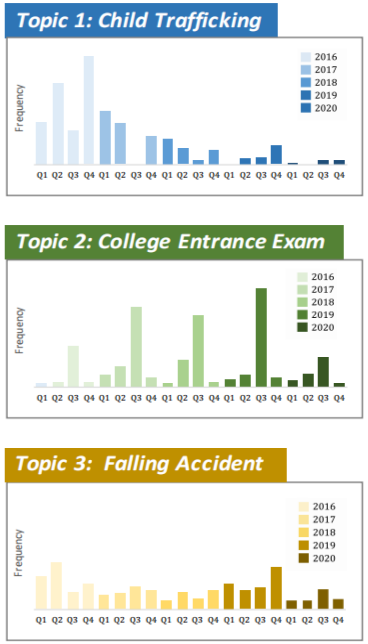
Text: Barcelona officially confirmed that Umtiti's COVID-19 test result was positive. As an asymptomatic patient he has begun home isolation. Umtiti did not follow the team to the UCL due to injury.

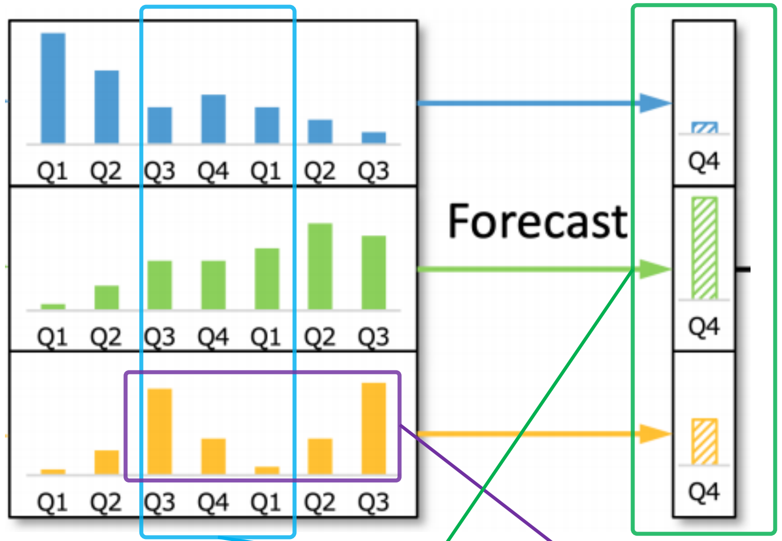
0

What are the paper going to achieve?

- Topic Distribution (especially of fake news) is dymanic, instable across times.
- ➢ Periodic? Seasonal? : The adjustment of the importance of news topics in training process, should be slightly prophetic according to the result of seasonal and periodic forecasting.
- ➢ Suddenly Appear or Suddenly Disappear?: Be susceptive to newly emerged and newly cool-down misinformation topics, then conduct fast adaption or forgetting for detectors.

➢ Methodology Overview





01

- Different topics of misinformation have diverse temporal patterns, regarding HEAT.
- Time series of misinformation topics have certain periodicity and seasonality.

0+

Primary Insight & Thought Process

02

If conduct positive-corelated instance reweighting on training process according to topic popularity, the model will be more adept at checking the veracity of instances on these topics.

03

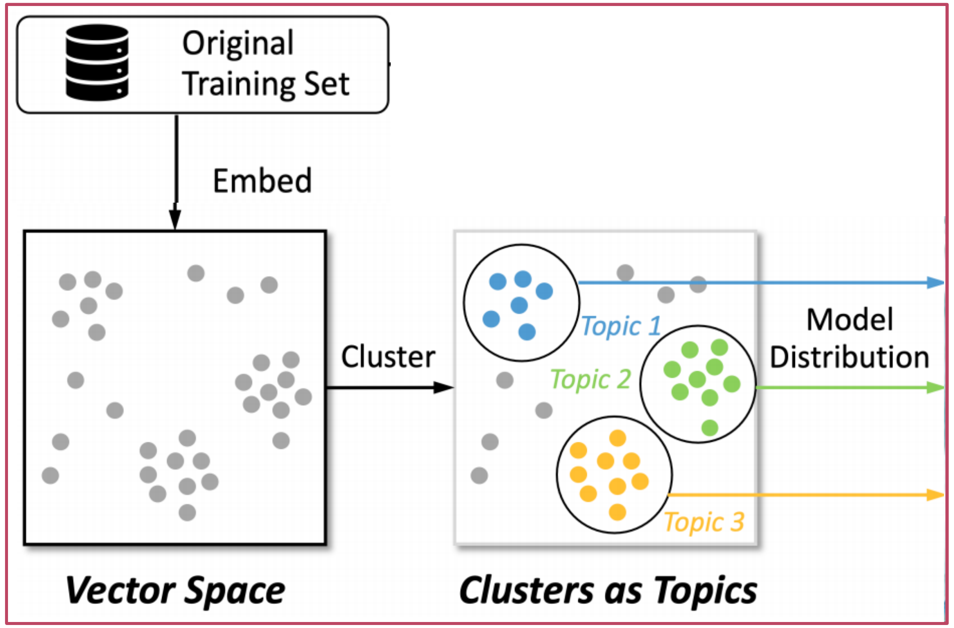
Generalize to Future: Modelling topic trends to forecast topic popularity distributions in the future, and aforehand do reweighing according to the future topic polularity, at the current time stamp

➢ Methodology (1): Topic Discovery

Adaptive Topic Discovery by Clustering

- **Why** employ **clustering**: **Lack prior knowledge** about the topic **number**, topic **semantics** and **definitions**.

- Employ **single-pass incremental clustering**:
  - ➢ No need to preset a cluster number. Totally adaptive.
  - ➢ **Support dymanic addition** of **new topics**. (**When** the **distance** of a new instance from all clusters **exceeds** a **threshold θsim=0.5**)



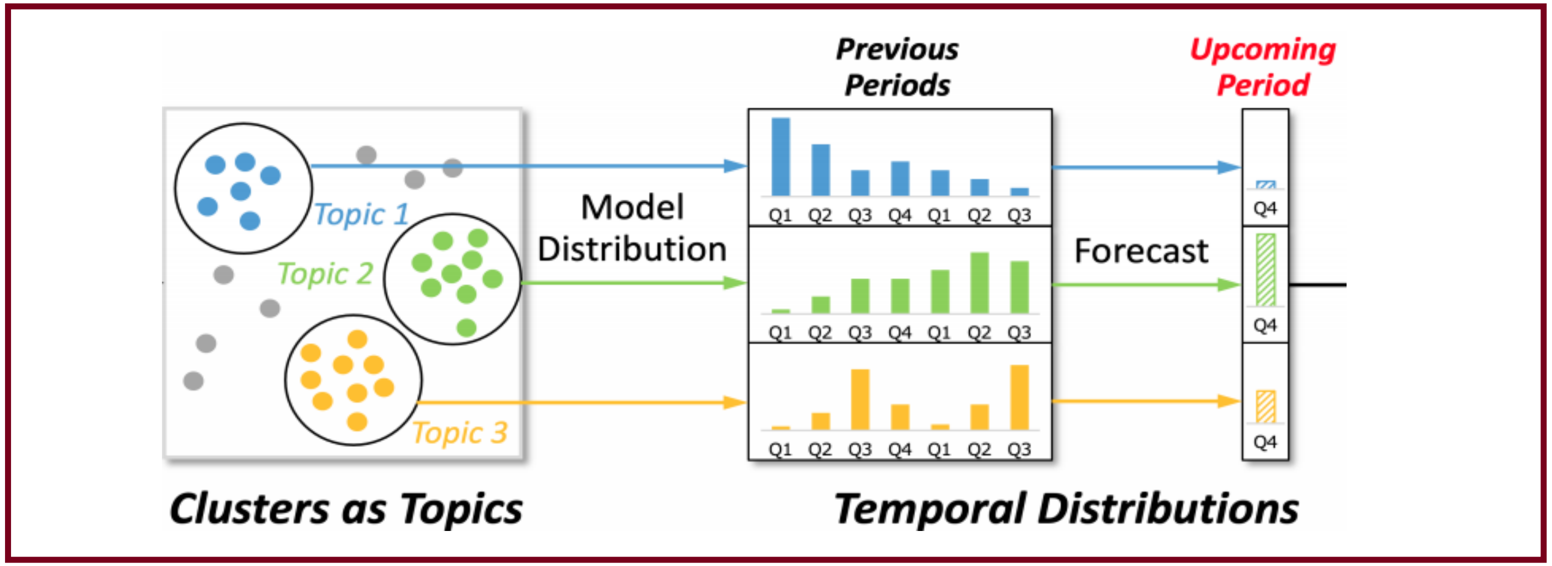
Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

➢ Methodology (2): Forecast Topic Trends



2

Approach: A Decomposable Time Series Model -- Prophet

- **A Linear trend with changepoint**: (to capture the regular trends)

$$g_i(f_{i,q}) = k_i f_{i,q} + m_i \qquad \text{where } k_i = k + \boldsymbol{a}(q)^T \boldsymbol{\delta} \qquad m_i = m + \boldsymbol{a}(q)^T \boldsymbol{\gamma}$$

$\boldsymbol{\delta}$, k and m are **learnable**

- Some **mistakes** in writting of the formulas above: **k and m are topic-specific**, not shared; quarter q is the independant variable; **f_(i, q)**, i.e., the number of news items within topic i in quarter q, is the **dependant variable of g_i( )** actually
- Calculate seasonal term using four univariation regressor, then do average summation. The **seasonal trend-popularity score** is written as: $s_i(f_{i,q})$
- Last, **combine** the two term, we can produce the **forecasting of topic trends p_i**:

$$p_i(f_{i,Q}) = g_i(f_{i,Q}) + s_i(f_{i,Q})$$

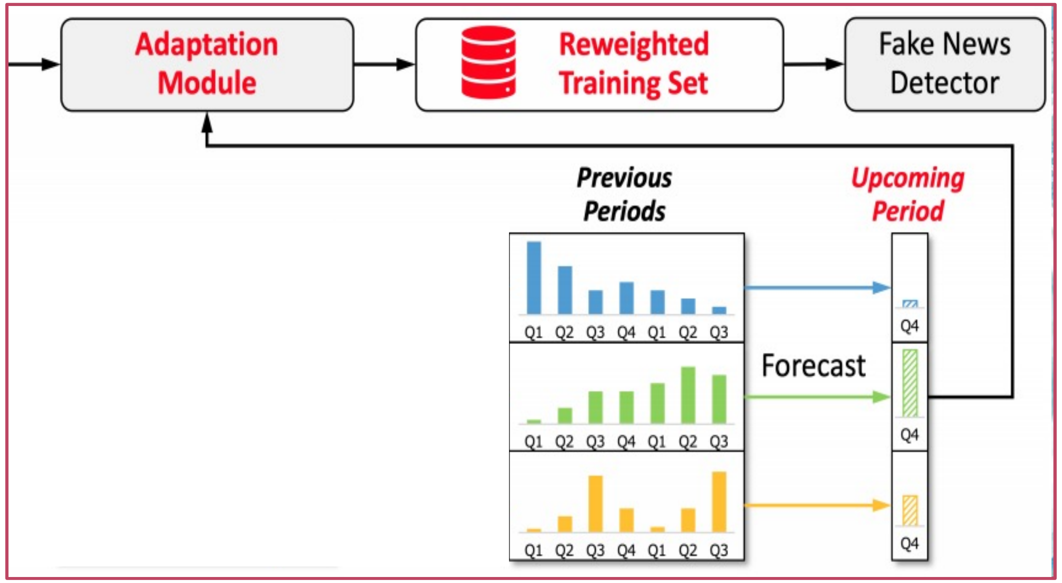
Methodology (3): Reweighting

Topic-level Reweighting

- **Filter out trivial topics** (without obvious regularity) using a **thershold** $\tilde{\theta}_{mape}$ on the **MAPE of prophet-fit.**
- **Topic-level** Reweighting: **news pieces** belonging to the **same topic share** the reweighting score:

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

- The **reweighing** score has an **upper bound** (>1) and a **lower bound** (<1).
- The score for **filtered-out** topics is **set to 1.**
- $w_{i,Q}$ **positively-corelated** to the **delta** between the **forecased and the current** $\frac{p_i(f_{i,Q})}{\sum_{i \in D_Q'} p_i(f_{i,Q})}$



◆ $w_{i,Q}$ **is to reweight the calculation of loss:**

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

✓ **It's also Model-Agnostic!!!**

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

➢ Comprehensive Evaluation

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**

3+

Experimental Result is Strong

- FTT outperforms the baselines in most cases. The average improvement of F1_fake is larger than that of F1real, suggestting FTT have super-better recall against fake news.
- Reasons: fake news often focuses on specific topics, so its topic distribution is more stable. And its trend is more periodic and seasonal.

➢ Extensive Analysis

Existed and New Topics on Test

- Intention: To analyze how FTT improves fake news detection performance.

➢ FTT sees **super-strong recall** capability for **newly added topics** on the **future** timestamp.

➢ **Precision** also get **improved**: the reweighting strategy in this paper can both focus on increasing trends and fading topics -- **more familiar with the past, more generalizable to the future**

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**

04

Future Challenge & Critique Summarization

Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

➢ Future Challenge: LLM-Generated Misinformation

Dataset	Human-written		Paraphrase Generation		Rewriting Generation		Open-ended Generation	
	No CoT	CoT	No CoT	CoT	No CoT	CoT	No CoT	CoT
ChatGPT-3.5-based Zero-shot Misinformation Detector								
Politifact	15.7	39.9	↓5.5 10.2	↓7.4 32.5	↓5.7 10.0	↓11.9 28.0	↓8.5 7.2	↓16.6 23.3
Gossipcop	2.7	19.9	↓0.4 2.3	↓2.2 17.7	↓0.5 2.2	↓2.7 17.2	↓0.1 2.6	↓1.0 18.9
CoAID	13.2	41.1	↓8.9 4.3	↓2.7 38.4	↓10.1 3.1	↓4.3 36.8	↓9.3 3.9	↓17.8 23.3
GPT-4-based Zero-shot Misinformation Detector								
Politifact	48.6	62.6	↓6.9 41.7	↓6.6 56.0	↓13.8 34.8	↓9.0 53.6	↓26.6 22.0	↓21.0 41.6
Gossipcop	3.8	26.3	↑0.8 4.6	↑3.7 30.0	↑1.5 5.3	↓1.3 25.0	↑1.3 5.1	↓0.6 25.7
CoAID	52.7	81.0	↓5.4 47.3	↑1.2 82.2	↓6.2 46.5	↓7.7 73.3	↓25.2 27.5	↓28.3 52.7
Llama2-7B-chat-based Zero-shot Misinformation Detector								
Politifact	44.4	47.4	↓12.2 32.2	↓9.6 37.8	↓16.3 28.1	↓19.6 27.8	↓25.5 18.9	↓25.2 22.2
Gossipcop	34.6	40.7	↑3.5 38.1	↓9.5 31.2	↓3.0 31.6	↓13.9 26.8	↓7.8 26.8	↓23.0 17.7
CoAID	19.8	23.3	↑4.6 24.4	↑15.1 38.4	↑1.1 20.9	↑15.1 38.4	↑15.1 34.9	↓4.7 18.6
Llama2-13B-chat-based Zero-shot Misinformation Detector								
Politifact	40.0	14.4	↓12.6 27.4	↓2.9 11.5	↓19.3 20.7	↓4.8 9.6	↓30.4 9.6	↓10.7 3.7
Gossipcop	10.8	7.8	↑3.9 14.7	↑4.8 12.6	↓0.8 10.0	↓2.2 5.6	↓2.1 8.7	↓0.9 6.9
CoAID	30.2	17.4	↑2.4 32.6	↓1.1 16.3	↓8.1 22.1	↓11.6 5.8	↓22.1 8.1	↓8.1 9.3

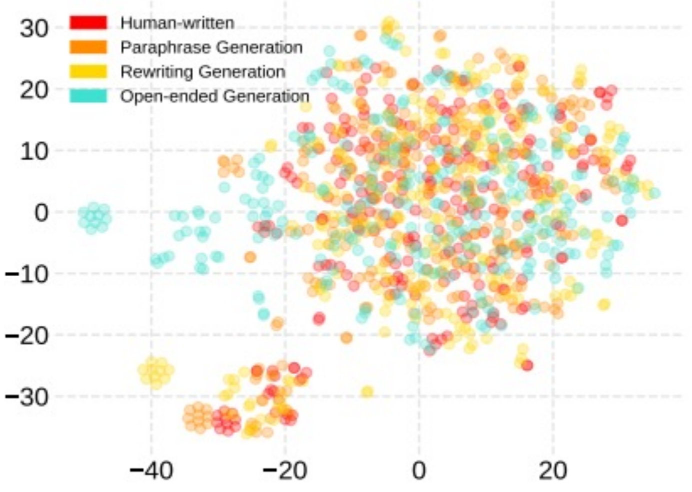


Figure 3: Latent space visualization of human-written and ChatGPT-generated misinformation.

- **LLM-generated Misinformation is in a new pattern: different semantic distributions and more difficult to be detected.**

- We **don't have enough** comprehensive **datasets** of **LLM-generated misinformation.**
- We don't know **whether LLMs** will **self-iterate** to bypass the detectors when generating misinformation

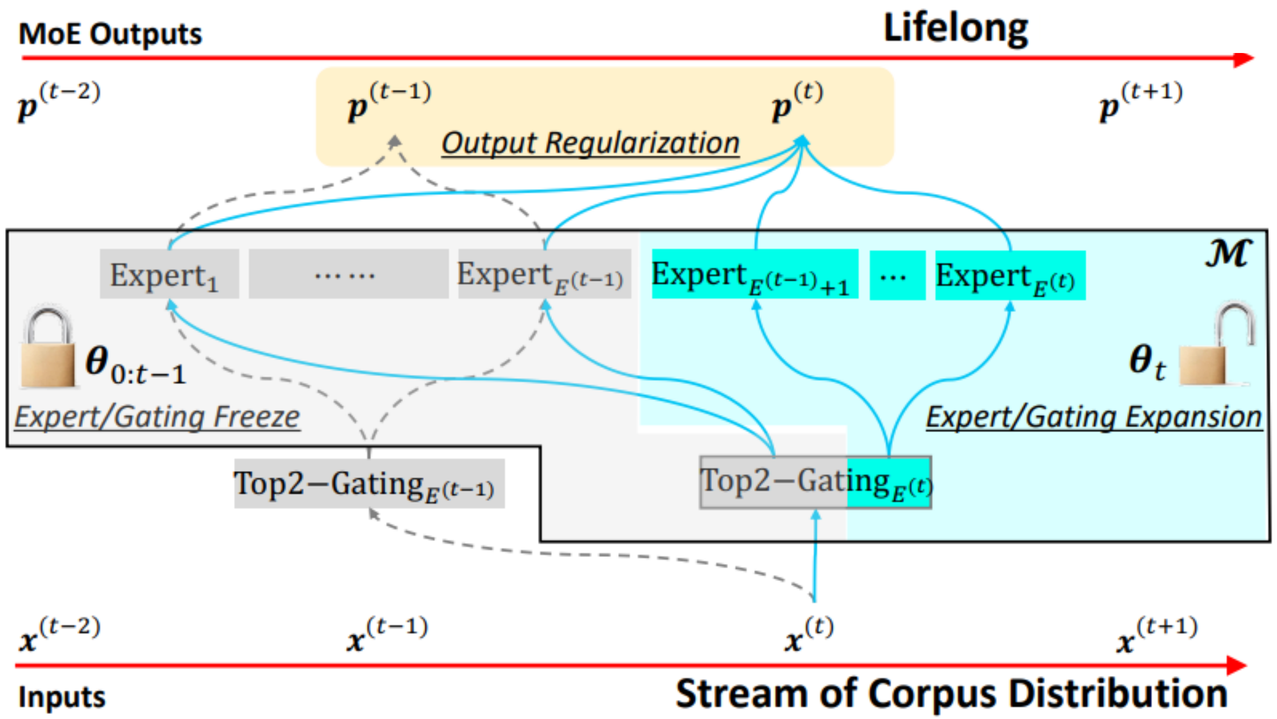
Recaps: Misinfo Det

Remove Entity Bias

Capture Topic Trends

Fut & Crit

➢ Future Challenge: Try to Unforget Anything



LAST

IF a detector can memorize any kinds of misinformation by a proposed lifelong training method......

➢ Critiques

Potential Shortcommings

- When the temporal interval is small, ENDEF might harm the performance. **Need to summarize applicable conditions.**

- The fittingness of the debiased model is not theoretically convinced.

◆ **Non-significant Imediate** Detection Capability: The FTT is **super effective** in the **2nd quarter** since the new topic appears; but not **theoritically better** than others in the **1st quarter**.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

THANKS

Geng. Zhao (趙 耕)