Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning

Qiang Sheng, Xueyao Zhang
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
shengqiang18z@ict.ac.cn
zhangxueyao19s@ict.ac.cn

Juan Cao
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
caojuan@ict.ac.cn

Lei Zhong
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
zhonglei18s@ict.ac.cn

CIKM'21

220303 Chia-Chun Ho

Outline

Introduction

Fake News Detection

- Fake news that spread on "online" social media continually cause "offline" real-world harms in crucial domain (politics, finance, public security).
- COVID-19 infodemic where thousands of fake news pieces spread through social media.
- Under such severe circumstances, developing fake news detection system has been critical for maintaining a trustful online news ecosystem.

Introduction

Way to detect fake news

- Proposed to extract hand-crafted features or deep-learning features.
  - From contents, social contexts, propagation networks, etc.
- In this paper,
  - The authors focus on the deep learning method based on textual contents.
  - Can be grouped as:
    - Pattern-based methods & Fact-based methods

Introduction

Pattern-based method

- Aim at learning shared features (patterns) among fake news posts and expect these features to generalize to unseen posts.

- Ideal model tends to predict the veracity melting more on the highly frequent use exclamation marks or the words that urge readers to repost (retweet).

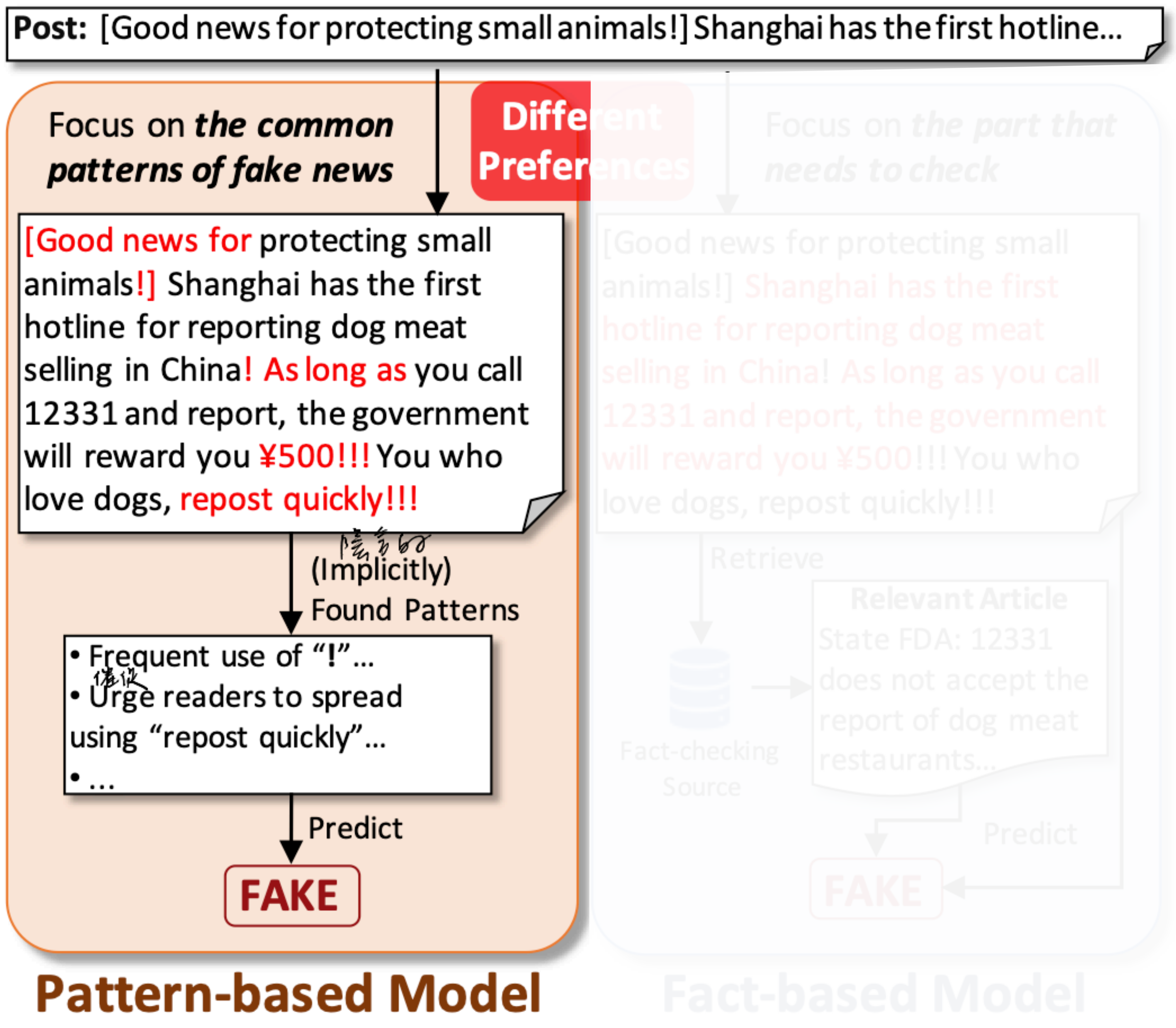


Figure 1: A motivating example. Ideally, given the same news post, the pattern-based and the fact-based model have *different preferences* on textual clues to predict whether the post is fake. The post is translated into English.

Introduction

Fact-based method

- Focus on the claim's veracity itself with the help from external fact-checking source.
- Ideal model retrieves to check whether the hotline accepts reports of dog meat selling.

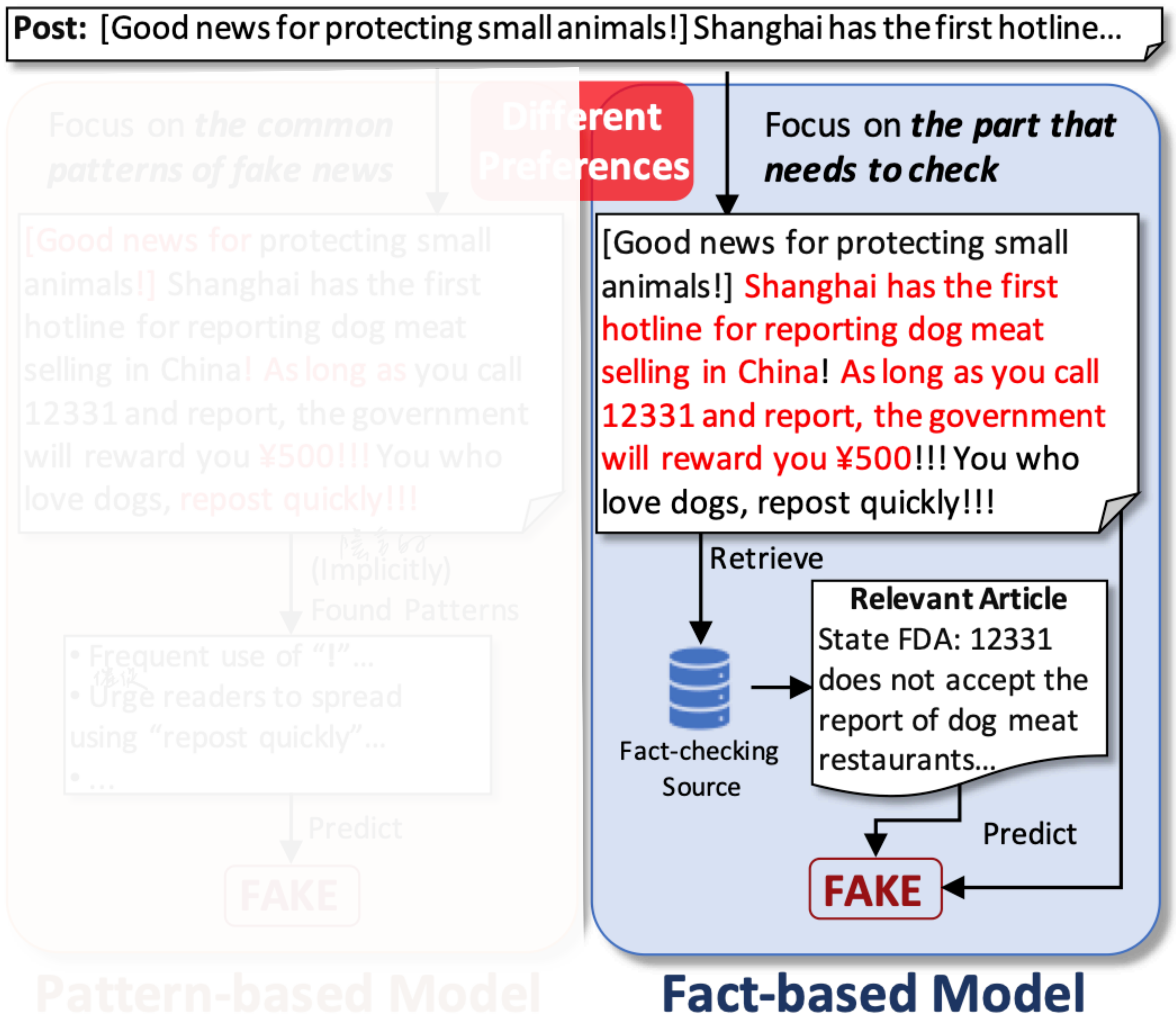


Figure 1: A motivating example. Ideally, given the same news post, the pattern-based and the fact-based model have *different preferences* on textual clues to predict whether the post is fake. The post is translated into English.

Introduction

Complementary methods

- The key difference between these two methods lies in their difference preferences of textual clues.
- See that the difference preferences of the two models lead to their complementary roles.
- Inspires to integrate pattern- and fact-based models with considering their preferences.

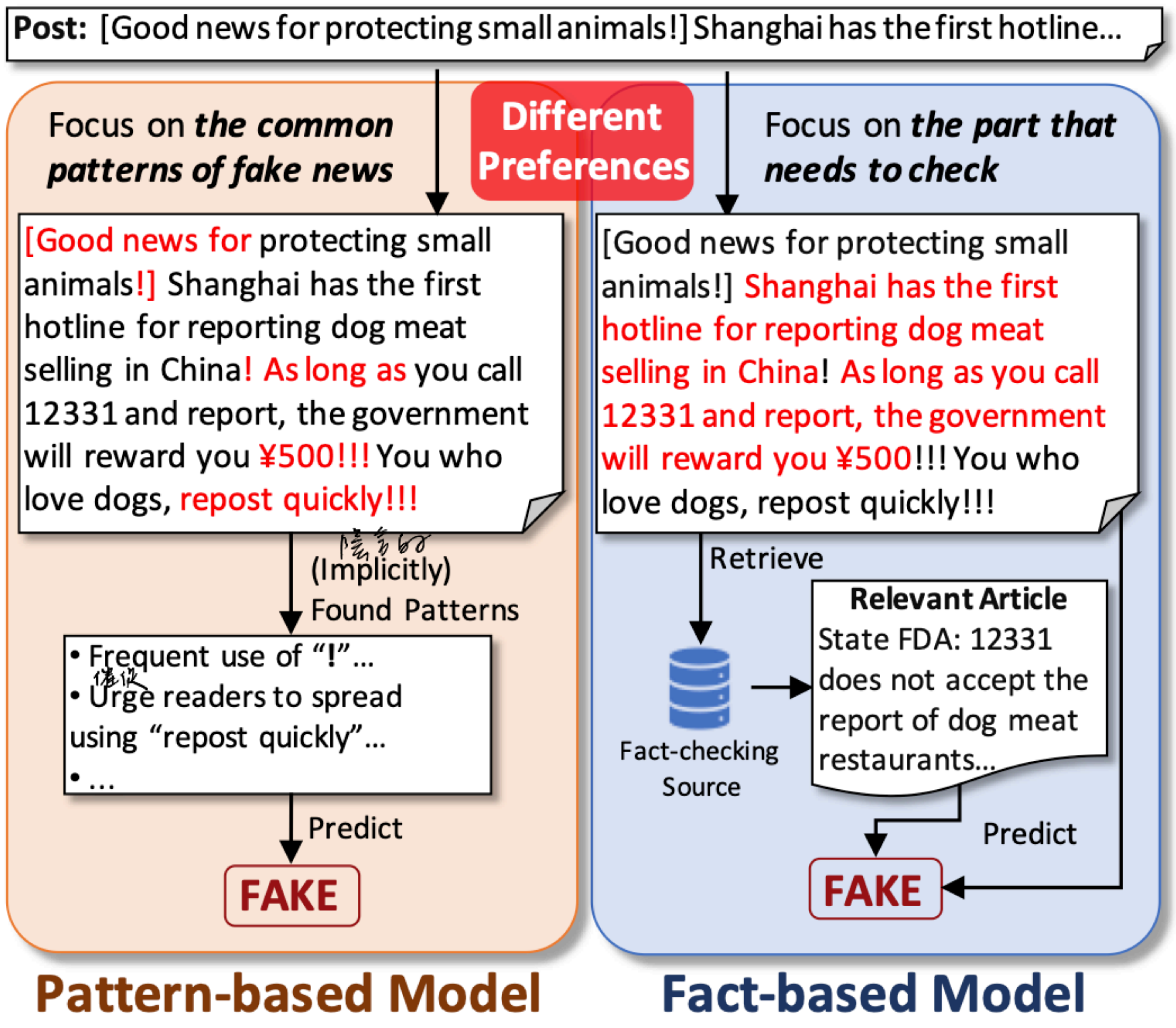


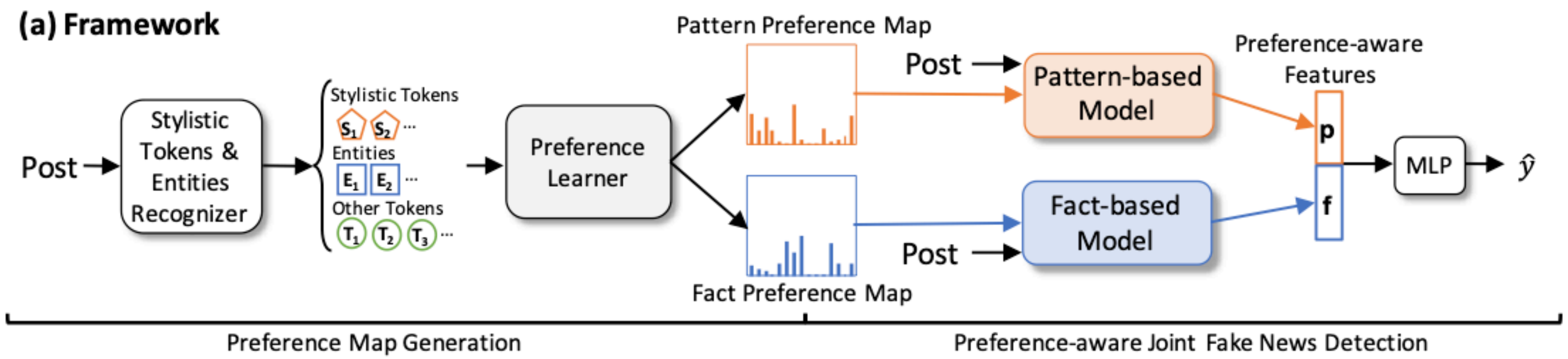
Figure 1: A motivating example. Ideally, given the same news post, the pattern-based and the fact-based model have *different preferences* on textual clues to predict whether the post is fake. The post is translated into English.

Introduction

Integrating two models into one framework

- The challenge lies in preference modeling:
  - The models, though having different preferences, generally lack the constraints to make themselves focus on preferred parts and ignore non-preferred parts of inputs.
  - Pattern-based model may overfit by memoizing frequently shown non-preferred words (e.g. event-specific words) in training set.
  - Fact-based model may be distracted from the part that describes a verifiable event.
- Preference of each model should be dynamically determined with contexts.
  - Making rule-based modeling inapplicable.

Introduction
Pref-FEND



- Proposed Preference-aware Fake News Detection (Pref-FEND) to learn the models' preferences simultaneously with joint fake news detection.
- Pref-FEND generates preference maps to assist each model to focus on its expected preferred part.
  - Recognize tokens (patterns, facts) in content and construct a heterogeneous graph and design a Heterogeneous dynamic GCN (HetDGCN) for node correlation learning.
  - The final correlation matrix is used by two preference-aware readout functions to generate the Fact and the Pattern Preference Map, respectively.

Introduction

Preference-aware Fake News Detection

- Beside the normal classification loss, the authors design two auxiliary losses as enhancements.
  - Minimize the similarity between the two maps.
  - Classification loss when the input maps are exchanged & ground-truth labels are reserved.
- Experiment results on two real-world datasets show that proposed Pref-FEND can effectively learn the models' preferences and improve the performance of both single preference (pattern/fact) and integrated models.

Introduction

Contributions

- To best of the authors' knowledge, Pref-FEND is the first that combines pattern- & fact-based fake news detection.
  - Discuss their complementary roles in FND and propose to consider their preferences for better integration.
- Pref-FEND leverages a heterogeneous dynamic GCN to learn model preferences and effectively integrates them for FND.
- Experiment results demonstrate the effectiveness of Pref-FEND on learning models' preferences and improving the detection performance for both single-preference model and integrated models.

Related Works

Pattern-based Fake News Detection

- Focus on writing styles.
  - Inject subjectivity, psycholinguistic, and moral foundations features into CNNs and RNNs.
- Some works attempt to differentiate the patterns across multiple topical categories.
- Recent trend of pattern-based methods
  - Refocus on the sentiment and emotional patterns, as the use of eye-catching terms in deceptive and fake post may manipulate the readers' emotions.

Related Works
Fact-based Fake News Detection

- Judge the veracity of a news piece (claim) more objectively, with references to pre-constructed external resources.
  - e.g. knowledge graphs, online encyclopedias, and scientific articles.
- Directly use articles retrieved by search engines as evidence to predict the news veracity.
- Use post-specific attention to model the post-article interactions.
  - Consider text entailment, such coherence and conflicts using attention mechanism.

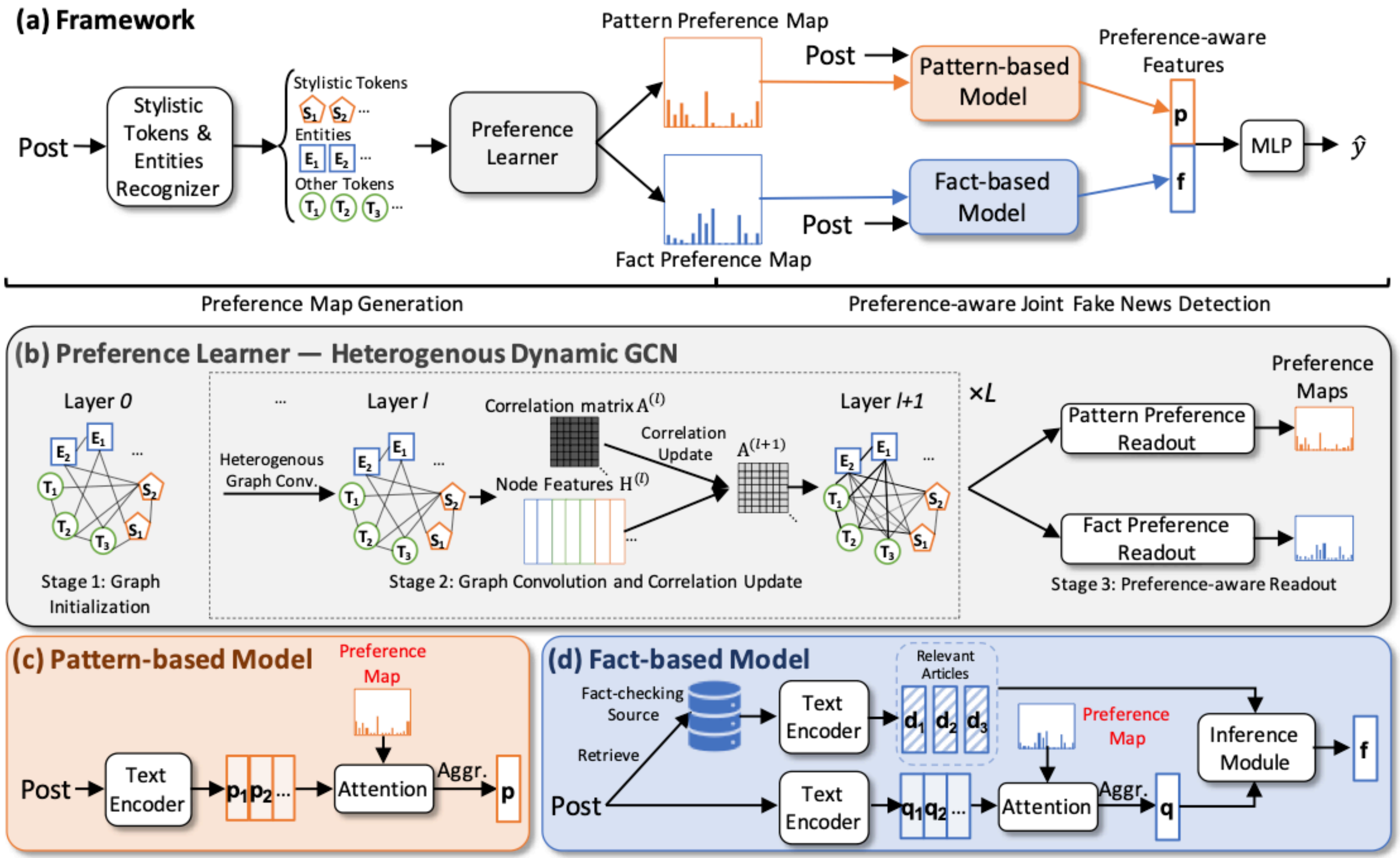
Related Works

Diff from proposed method

- In this paper, Pref-FEND do not develop better pattern- or fact-based methods.
- But integrate the existing ones for comprehensively detecting fake news on texts.

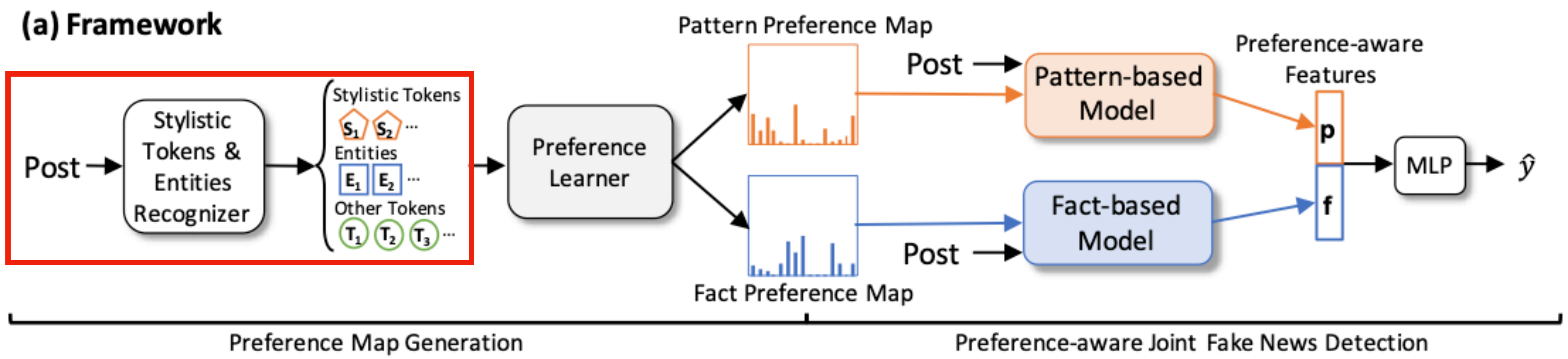
Methodology

Pref-FEND



Methodology

Preference Map Generation

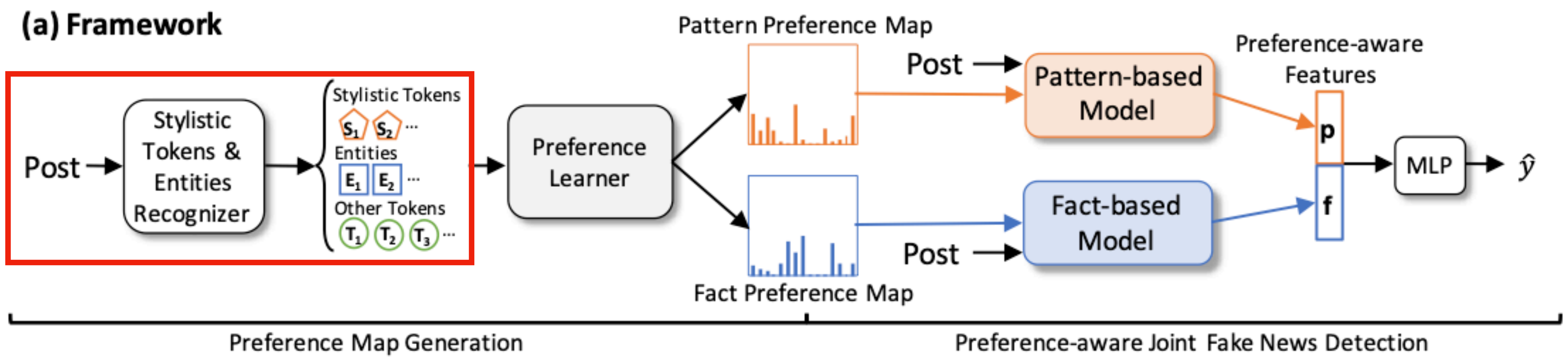


- Assume that post has n tokens, a preference map is a score distribution of length n where i-th token is preferred by corresponding fake news detection model.

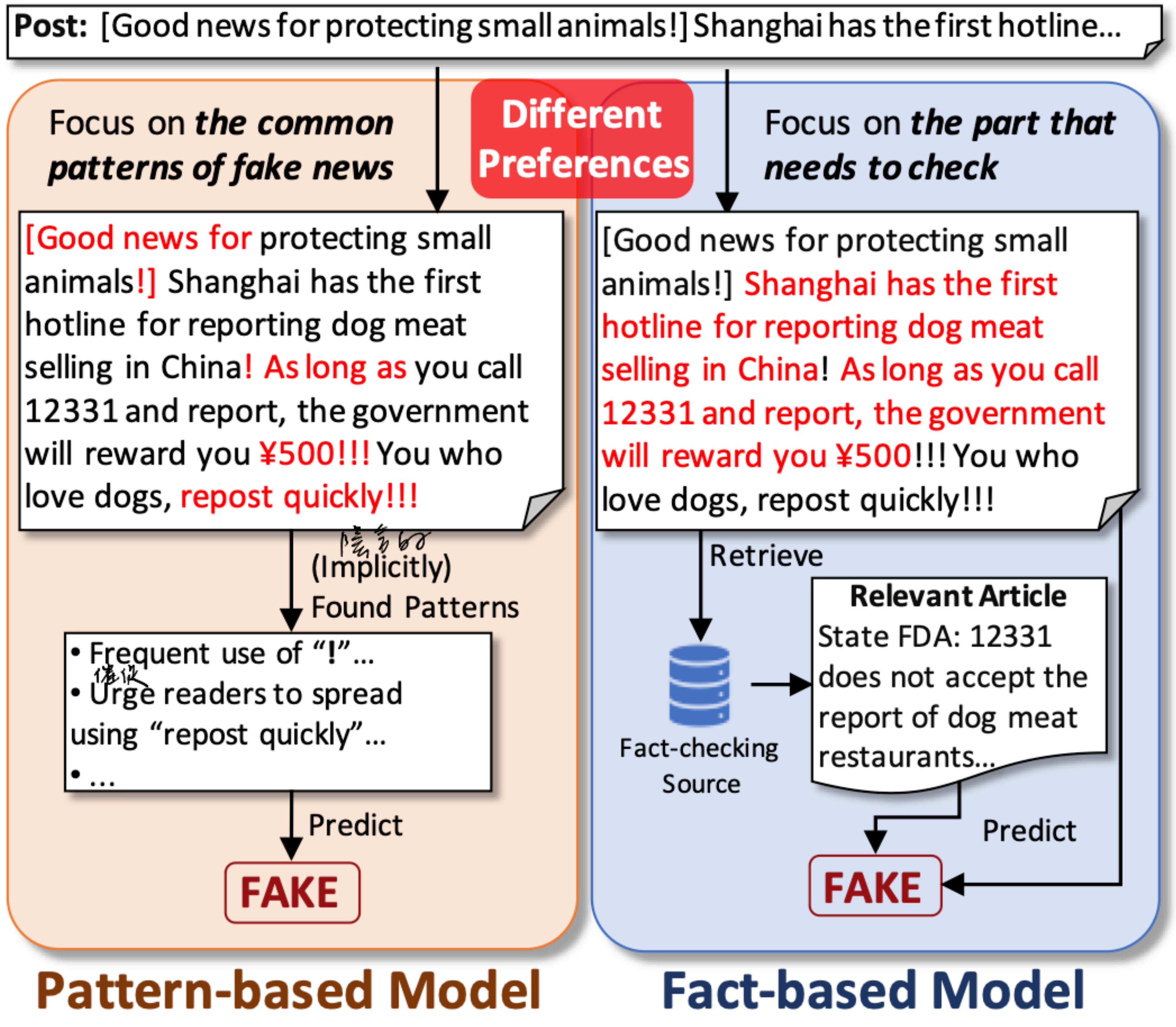
- Generate Pattern Preference Map and Fact Preference Map.
  - $\mathrm{m_P} = [\mathrm{m_{P}}_i]_{i=1}^{n}, \ \mathrm{m_F} = [\mathrm{m_{F}}_i]_{i=1}^{n}$
  - All scores are in [0,1].
  - Sum of each map is 1.

Methodology

Tokens recognition

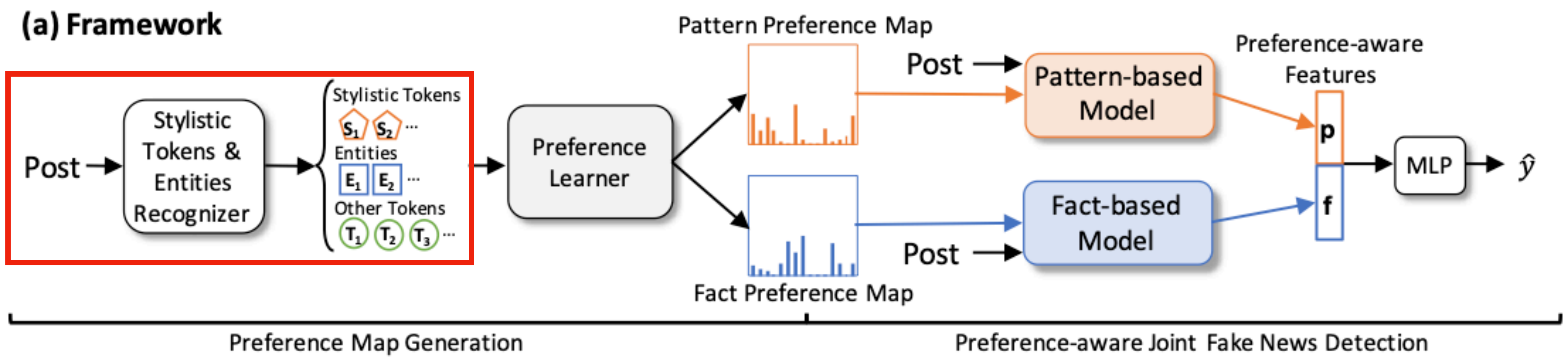


- As illustrated in previous:
  - Pattern-based model focuses on common patterns (styles).
  - Fact-based one focuses on verifiable objective claims.
- Recognize tokens that are likely to represent writing styles or key objective elements.



Methodology

Tokens recognition

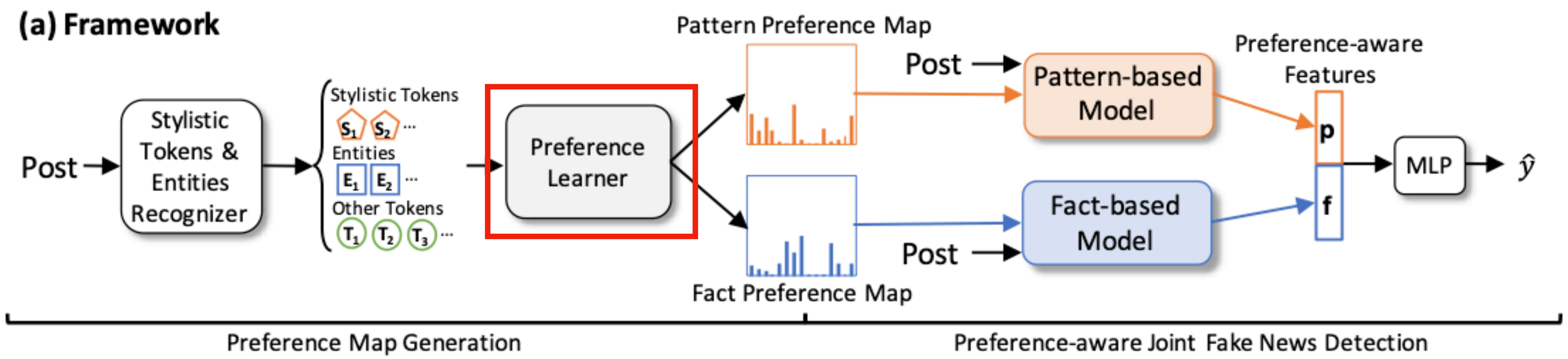


- To indicate patterns, recognize a set of stylistic tokens $S = \{s_1, \cdots, s_{n_s}\}$
  - e.g. emotional words, pronouns, punctuation marks
- To indicate facts, extract the entities $E = \{e_1, \cdots, e_{n_e}\}$.
- These indicating tokens are derived using pre-constructed dictionaries and public tools.
- To token excluded by S and E are in a set $T = \{t_1, \cdots, t_{n_t}\}$.
  - Where $n_t = n - n_s - n_e$

Type	For Weibo	For Twitter
Negation Word	HowNet Bilingual Dictionary [9]	
Degree Word		
Sentiment Word		
Proposition Word		
Punctuation	[64]	
Pronoun		
Emoticon	List of Emoticons [55] [64]	
Emotional Ontology	Affective Lexicon Ontology [60]	NEC Emotion Lexicon [29]

Methodology

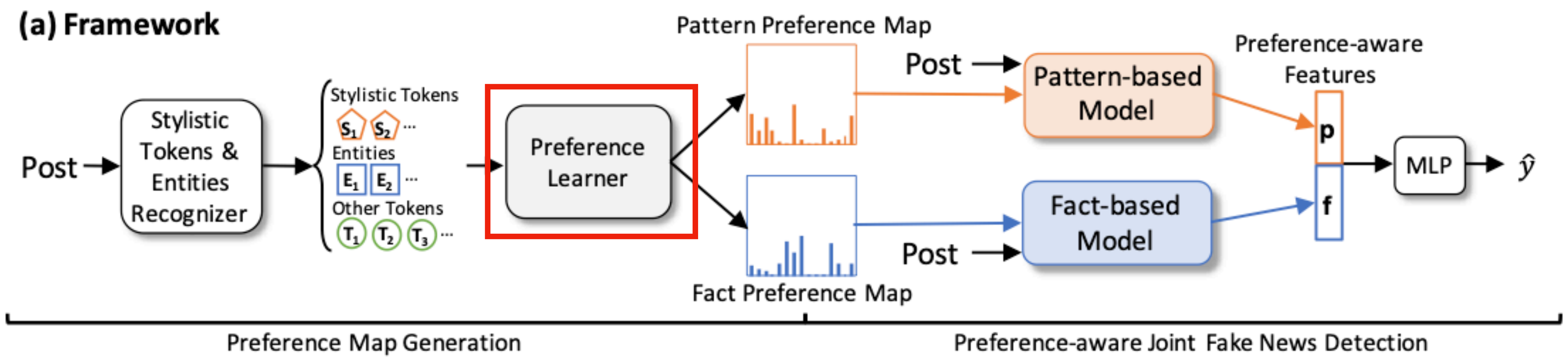
Heterogeneous DGCN



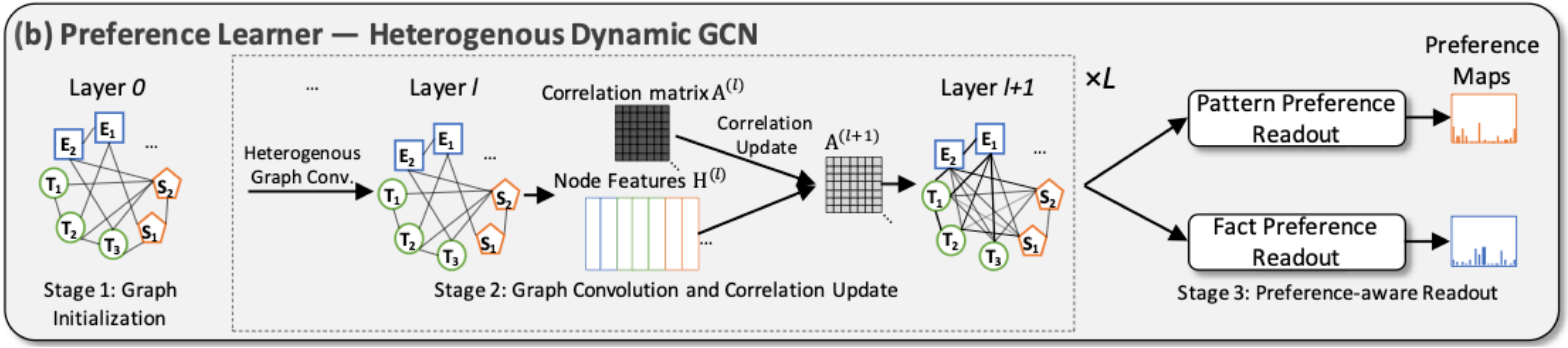
- Although the tokens & entities recognized by tools provide a good prior to what token might be preferred.
  - Directly using the recognition result for map generation is insufficient.
- The coverage is limit.
  - Leading the map to overlook some other preferred and useful tokens for detection models.
- A token's preference score should be dynamically and sufficiently interact with each other.

Methodology

Heterogeneous DGCN

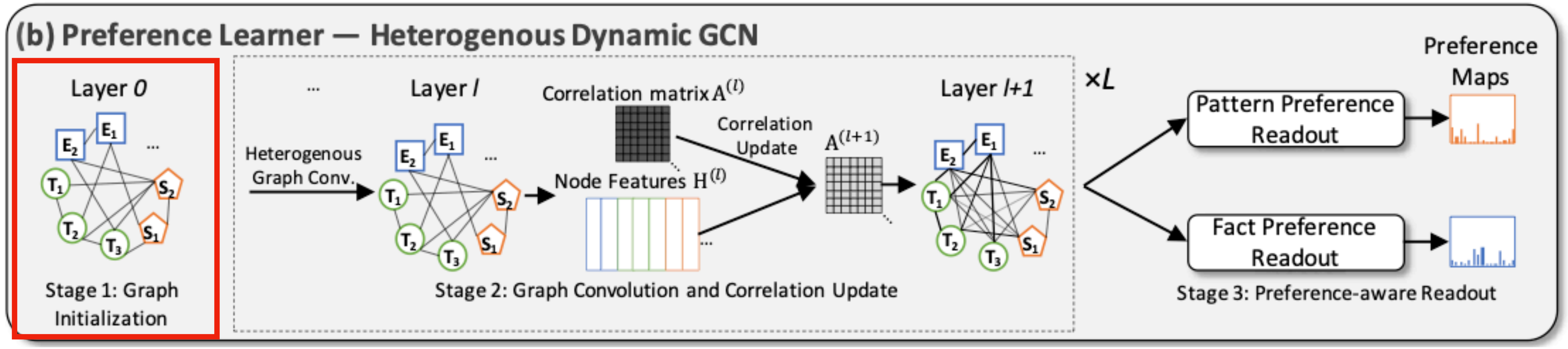


- Design a graph-based preference learner, HetDGCN.



Methodology

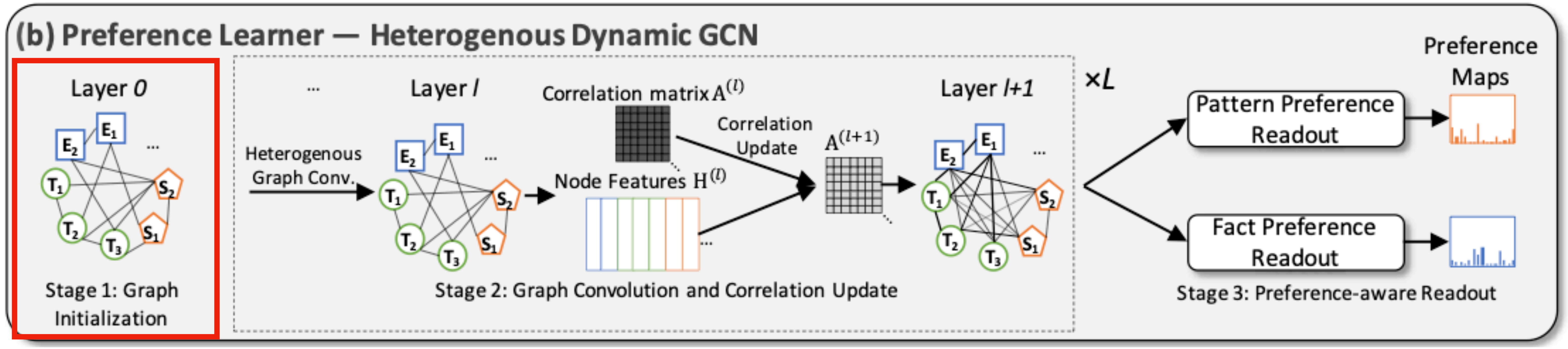
Graph Initialization



- Construct a heterogeneous graph G.
  - Nodes $\in \{S, E, T\}$, edge represents the correlation between the connected tokens.
  - Node representation is initialized with pre-trained language model (BERT).
  - $\mathbf{H}^{(0)} = [\mathbf{H}_S^{(0)}, \mathbf{H}_E^{(0)}, \mathbf{H}_T^{(0)}] \in \mathbb{R}^{n \times d}$
  - Edge weights are initialized with calculating the cosine similarity of token pair.
  - $\mathbf{A}^{(0)}(i,j) = \dfrac{\mathbf{h}_i^{(0)} \cdot \mathbf{h}_j^{(0)}}{2\|\mathbf{h}_i^{(0)}\|\|\mathbf{h}_j^{(0)}\|} + 0.5$, $\mathbf{h}_i^{(0)}, \mathbf{h}_j^{(0)}$: initial node features.

Methodology

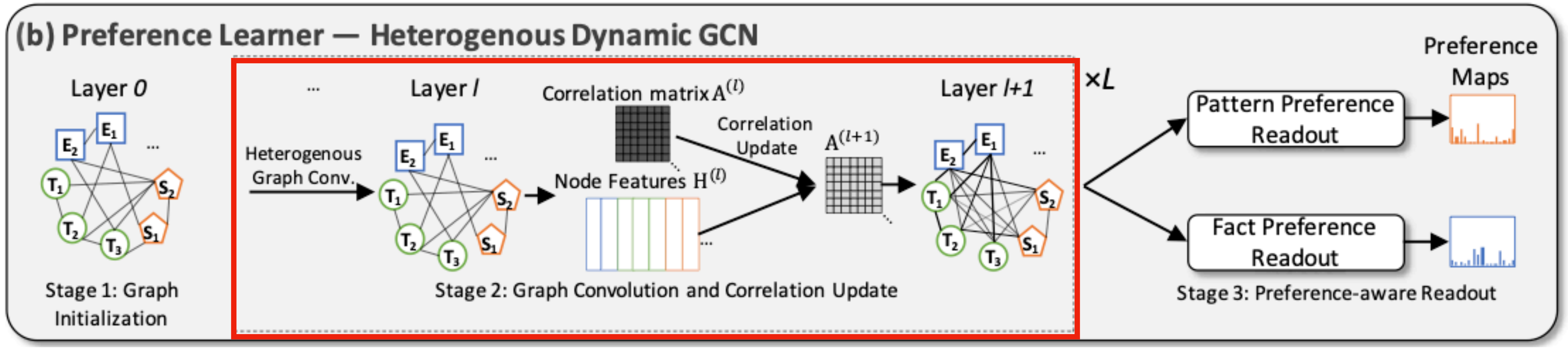
Graph Initialization



- Define the normalized correlation matrix of the l-th layer
  - $\hat{\mathbf{A}}^{(l)} = \left(\mathbf{D}^{(l)}\right)^{-\frac{1}{2}} \mathbf{A}^{(l)} \left(\mathbf{D}^{(l)}\right)^{-\frac{1}{2}}$
  - $\mathbf{D}^{(l)}$: degree matrix of the l-th layer where $\mathbf{D}^{(l)}(i,i) = \sum_j \mathbf{A}^{(l)}(i,j)$

Methodology

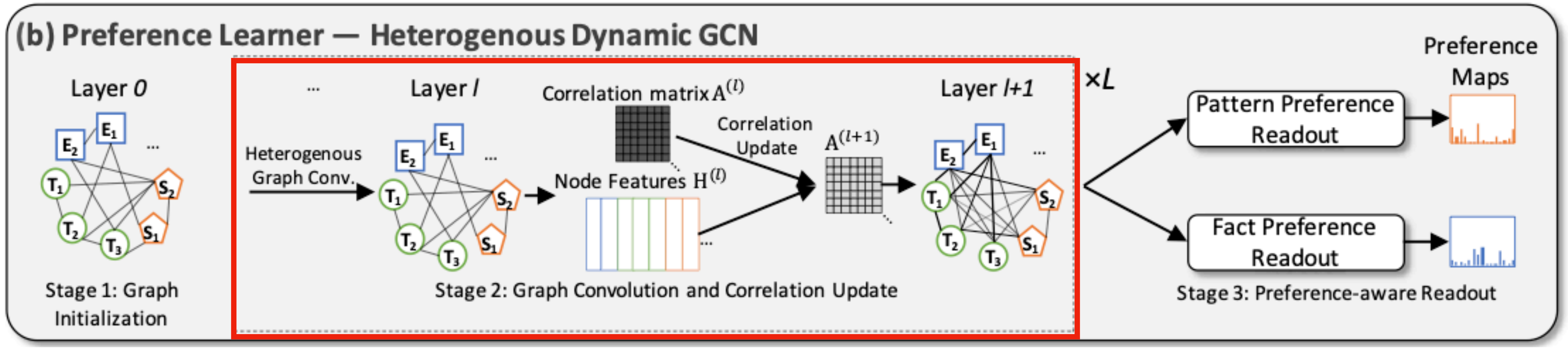
Graph Convolution



- Different types of nodes describe different aspects of the given text which expect to distinguish for preference learning.
- Instead of using standard GCN for node interaction, use a heterogenous graph convolution.
    - Separately handle the neighbors of different types and then aggregate the interacted features.
- Further, use a dynamic correlation matrix which is updated each layer according to the present node similarity and expect the final correlations could reflect the bias of the nodes in the context.

Methodology

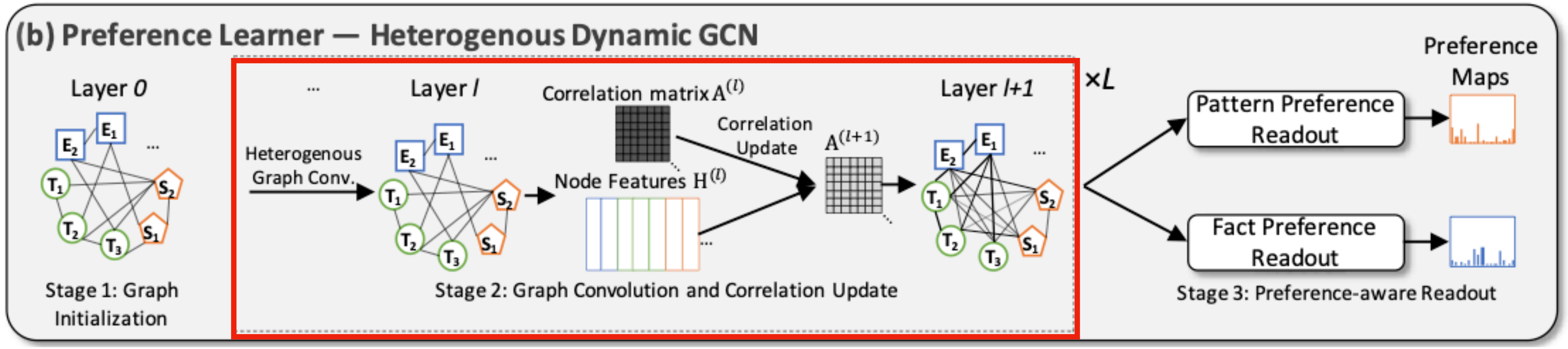
Graph Convolution



- In detail, the feature matrix of $(l+1)$-th layer is calculated with
  - $$\mathbf{H}^{(l+1)} = \mathrm{ReLU}\left(\sum_{\tau\in\mathcal{T}} \hat{\mathbf{A}}_{\tau}^{(l)}\mathbf{H}_{\tau}^{(l)}\mathbf{W}_{\tau}^{(l)}\right)$$
  - $\hat{\mathbf{A}}_{\tau}^{(l)}$: sub-matrix of the correlation matrix of the l-th layer $\hat{\mathbf{A}}^{(l)}$
  - $\hat{\mathbf{W}}_{\tau}^{(l)}$: learnable weight matrix of the type τ in this layer.

Methodology

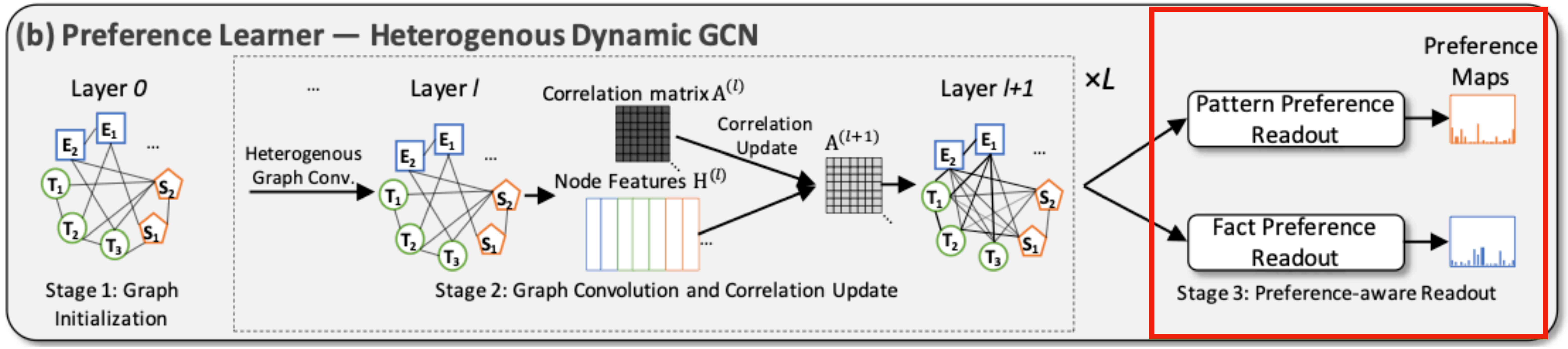
Correlation update



- Then, the correlation matrix is updated using
  - $\Delta \mathbf{A}^{(l+1)} = \sigma(\mathbf{H}^{(l+1)}\mathbf{W}_A^{(l+1)}\mathbf{H}^{(l+1)T})$
  - $\mathbf{A}^{(l+1)} = \alpha\mathbf{A}^{(l)} + (1-\alpha)\Delta\mathbf{A}^{(l+1)}$
  - $\mathbf{W}_A^{(l+1)}$: learnable weight matrix for update correlations
  - σ: sigmoid function, α: trade-off factor in [0, 1]

Methodology

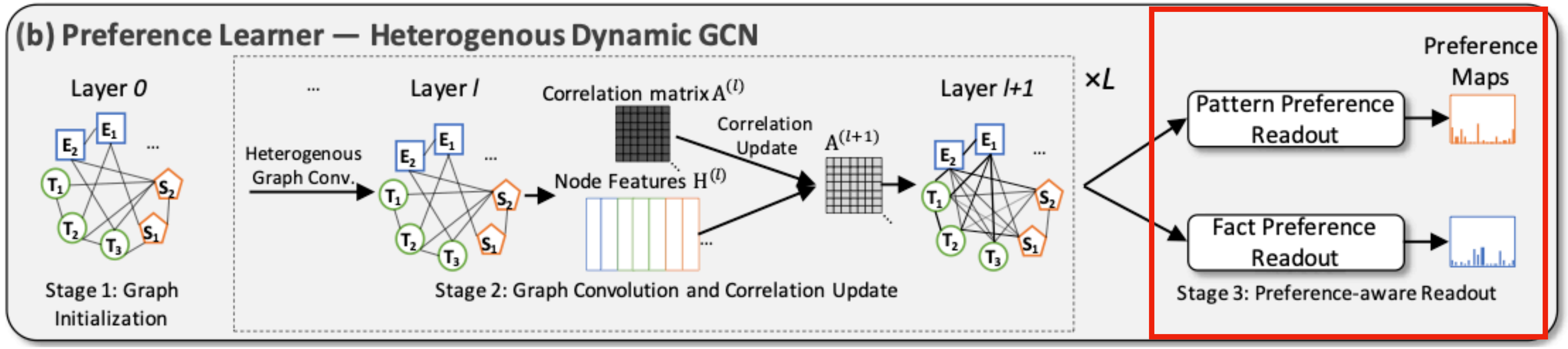
Preference-aware Readout



- After the L-layer HetDGCN, obtain the correlation matrix $\mathbf{A}^{(L)}$.
- Expect to estimate the preference levels to pattern- & fact-based models of each token.
- For the i-th node, the pattern preference score m_{Pi} is calculated by its correlation with any nodes except those representing entity token:
- $$\mathrm{m}_{\mathrm{P}i} = \sum_{j=1}^{n} \mathrm{A}^{(L)}(i,j) - \sum_{k=1}^{n_e} \mathrm{A}_E^{(L)}(i,k)$$

Methodology

Preference-aware Readout



- Similarly, the fact preference score excludes the correlation with the stylistic nodes:

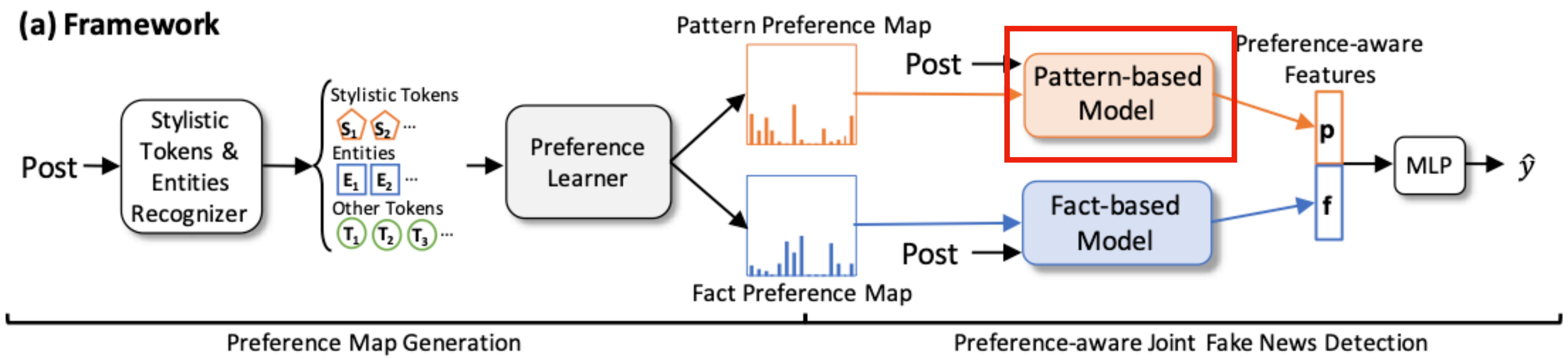
- $$m_{Fi} = \sum_{j=1}^{n} A^{(L)}(i,j) - \sum_{k=1}^{n_s} A_S^{(L)}(i,k)$$

- Finally, the preference maps are obtained by normalized the correlation sums of each token:

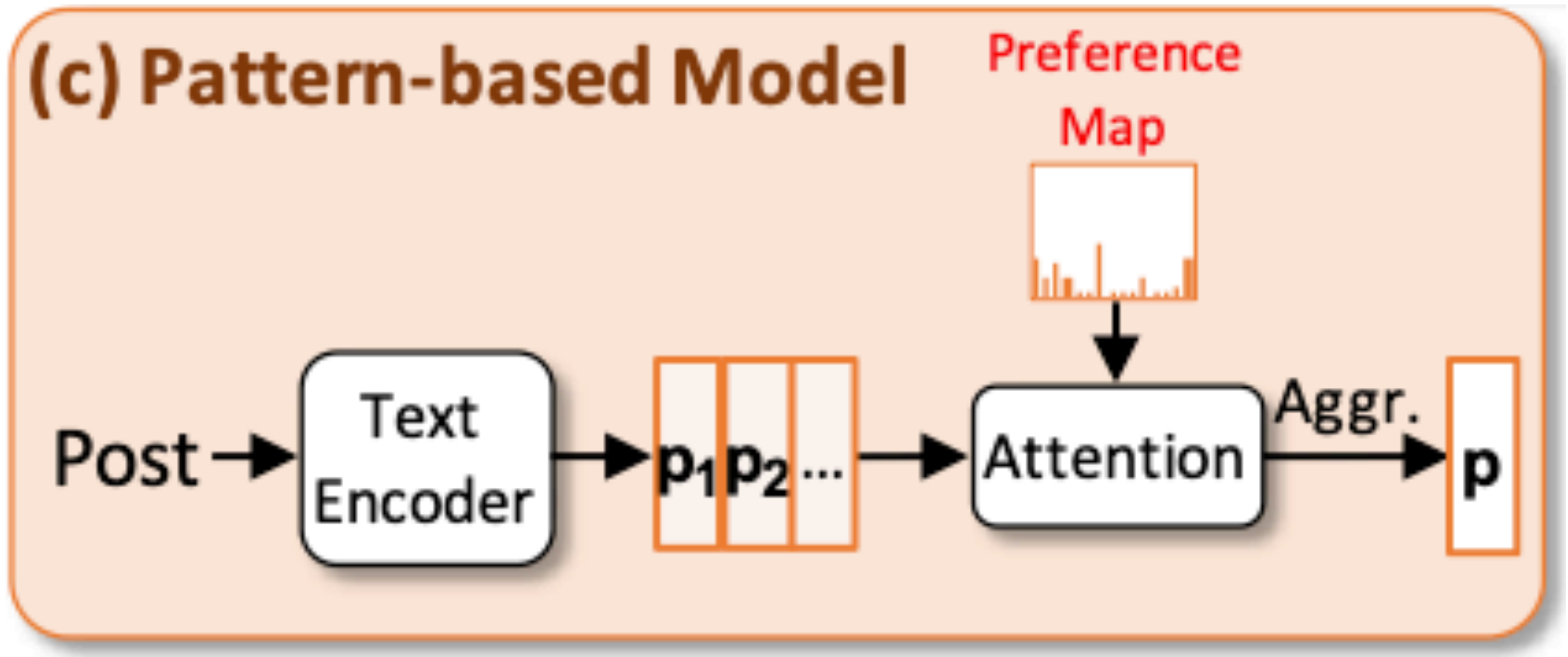
- $$m_P = \left[\frac{m_{Pi}}{\sum_j m_{Pj}}\right]_{i=1}^{n}, \; m_F = \left[\frac{m_{Fi}}{\sum_j m_{Fj}}\right]_{i=1}^{n}$$

Methodology

Pattern-based model

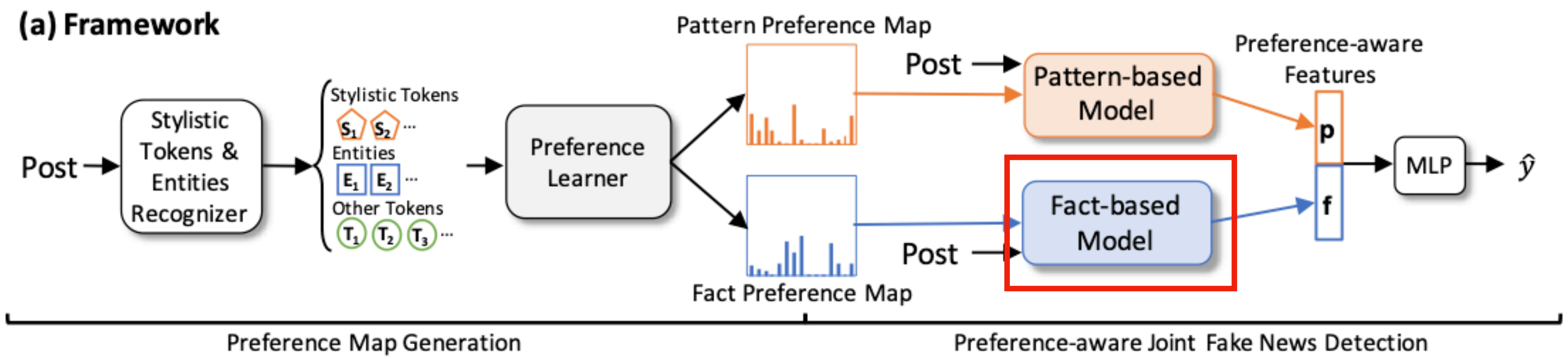


- A typical pattern-based uses a textual extractor to obtain a vector for final prediction.
- Here use the Patter Preference Map as attention weights to make the model attend to its preferred tokens in the post P.
- For example, if the extractor is a BERT whose output is $[p_1, \cdots, p_n]$.
- Aggregated vector is calculated as
  - $$\mathrm{p} = \sum_{i=1}^{n} \mathrm{m}_{\mathrm{P}_i} \mathrm{p}_i$$

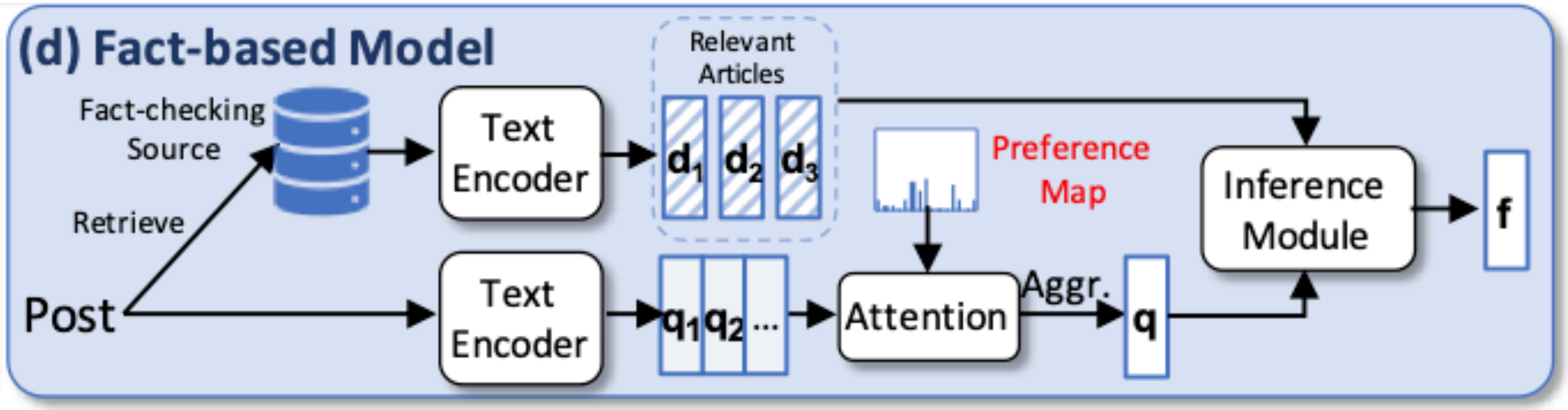


Methodology

Fact-based model

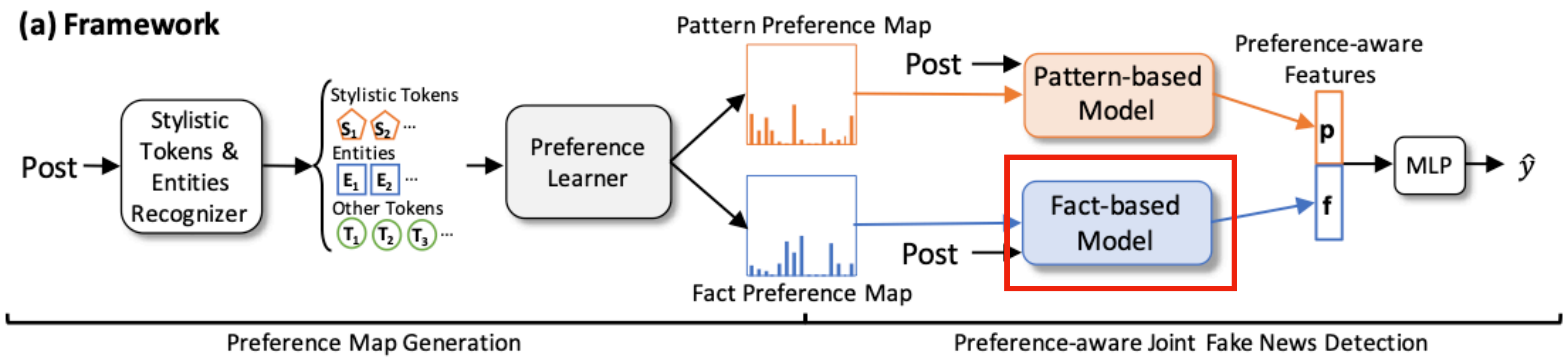


- In a typical fact-based model, the post P are first used to retrieve from a fact-checking source to collect the related articles (or, evidence) D.
- Assuming n_f articles are returned, represent the articles in D as $[d_1; \cdots; d_{n_f}]$.
- Then the post and evidence vectors are fed into an inference module, which is often designed to capture the complicate interactions such as coherence and conflicts between P, D.



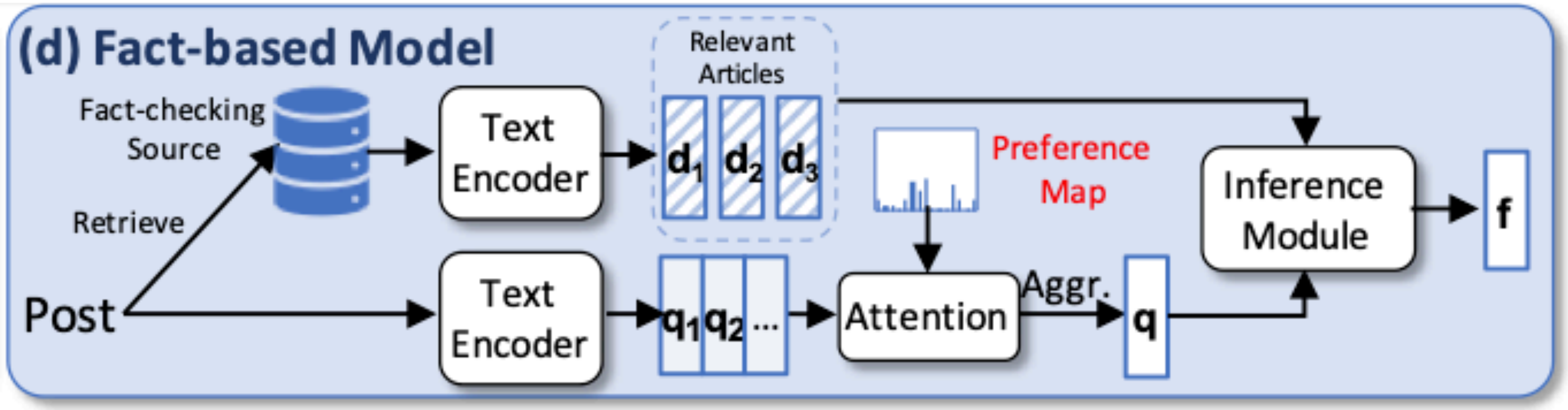
Methodology

Fact-based model



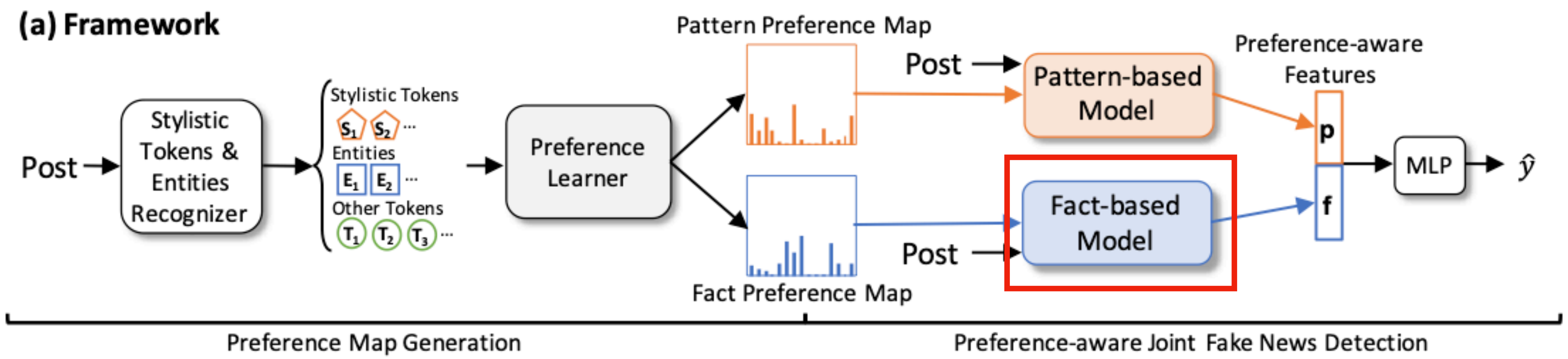
- The output vectors of inference module $\mathbf{f}$, which implicitly represent the relationship pf the post-evidence pairs, used for final prediction.

- To avoid the inference of non-check-worthy parts, the Fact Preference Map guides the inference module by using attention mechanism to aggregate the token vectors in P before post-evidence inference.



Methodology

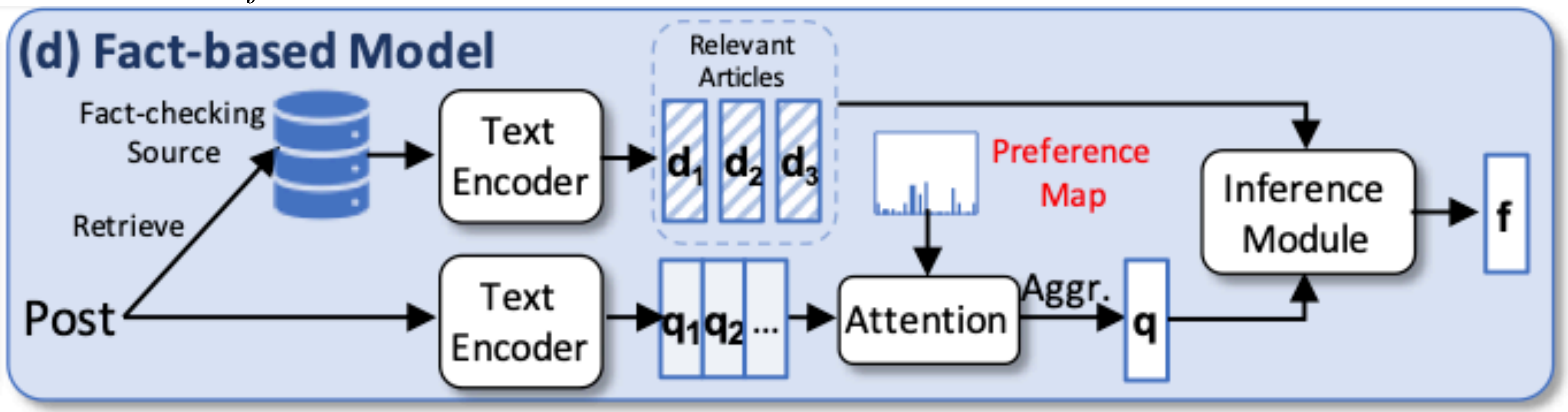
Fact-based model



- The final vector is calculated as

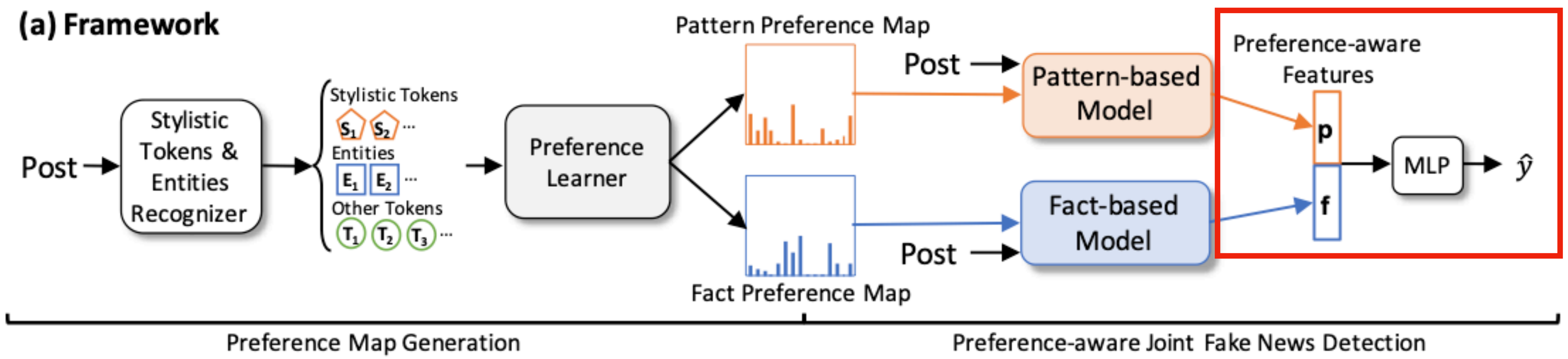
- $$q = \sum_{i=1}^{n} m_{Fi} q_i$$

- $$\mathbf{f} = \text{InferenceModule}(\mathbf{q}, [\mathbf{d}_1; \ldots; \mathbf{d}_{n_f}])$$



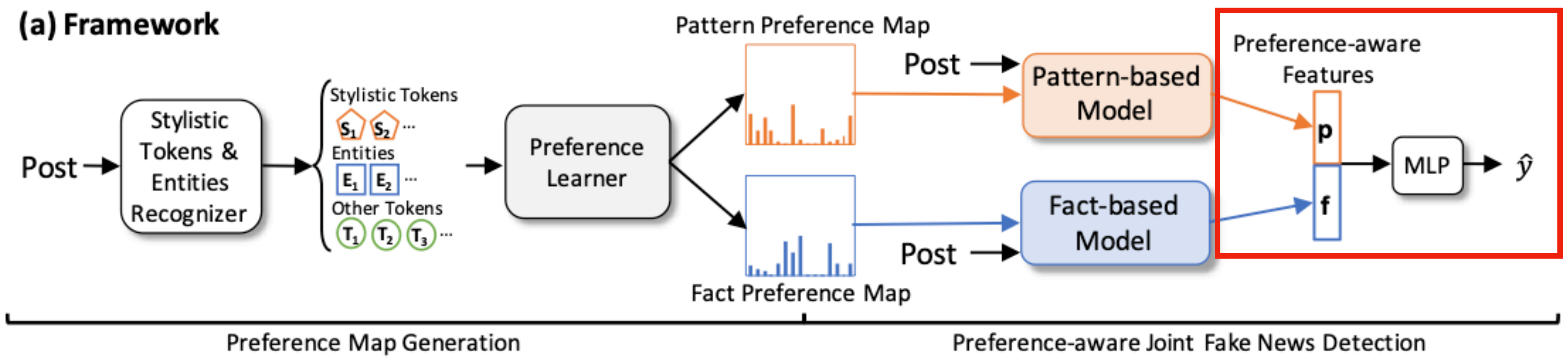
Methodology

Joint Detection



- For final prediction, concatenate the output vector of pattern- & fact-based models and feed it into a MLP and obtain the prediction $\hat{y}$:
  - $\hat{y} = \mathrm{MLP}([\mathbf{p};\mathbf{f}])$

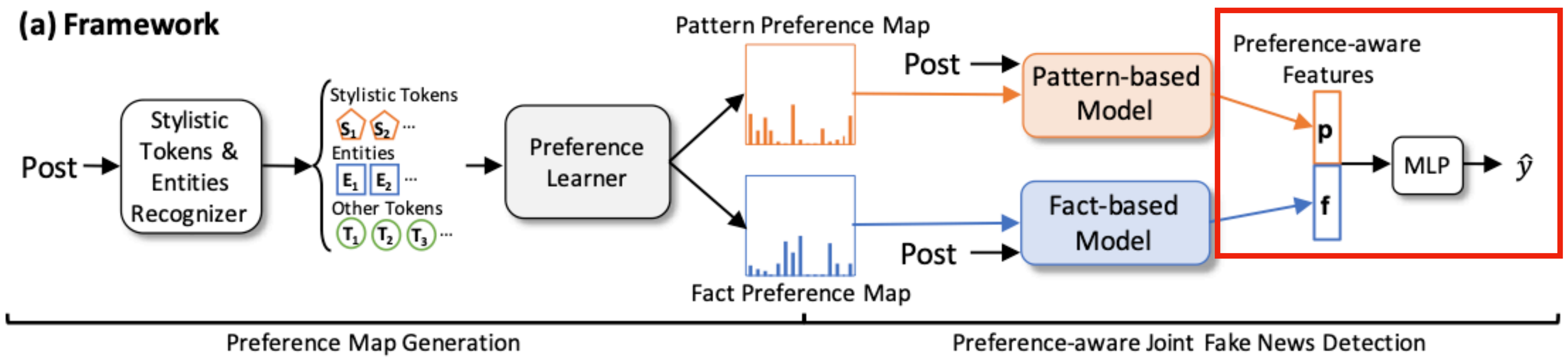
Methodology
Losses



- Use three losses to supervise:
  - Prediction of binary (fake/real) classification.
  - Differentiation of the two preference maps.
- For the first goal, minimize the cross-entropy loss between $\hat{y}, y$.
  - $\mathscr{L}_{cls}(y, \hat{y}) = \text{CELoss}(y, \hat{y})$
  - $\text{CELoss}(y, \hat{y}) = -y \log p - (1-y)\log(1-p)$

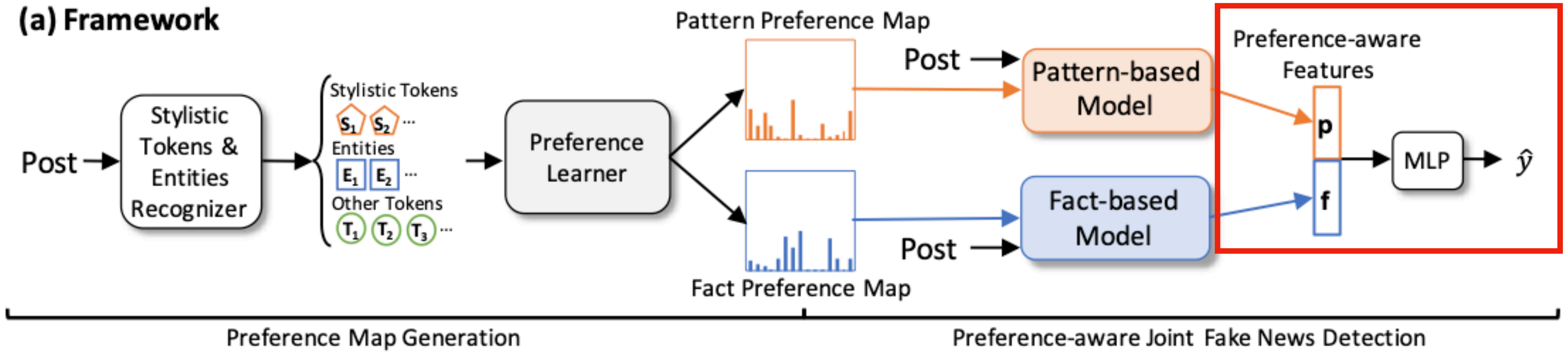
Methodology

Losses



- For second goal, consider the reciprocal roles of the two models and let them supervise mutually.
- Minimize the cosine similarity between the Pattern & the Fact Preference Map.
- $$\mathcal{L}_{\cos} = \frac{\mathrm{m_P} \cdot \mathrm{m_F}}{||\mathrm{m_P}||\,||\mathrm{m_F}||}$$

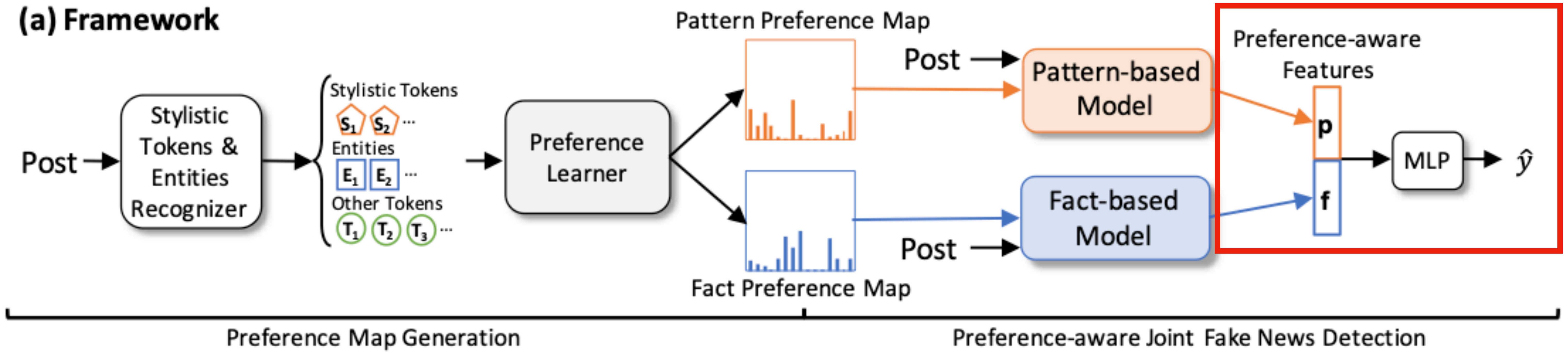
Methodology
Losses



- Cross-entropy loss under the condition that the input maps for the two models are exchanged and the ground-truth label is reserved.
- $\mathcal{L}_{cls}\left(y_{rev}, \hat{y}'\right) = \text{CELoss}\left(y_{rev}, \hat{y}'\right)$
- $y_{rev} = |1 - y|$
- $\hat{y}' = \text{MLP}([\mathbf{p}', \mathbf{f}'])$
- When receiving non-preferred information, the models are expected to be misled and generate non-distinctive features.

Methodology

Losses



- The total loss of a sample to minimize is
  - $\mathscr{L} = \beta_1 \mathscr{L}_{cls}(y, \hat{y}) + \beta_2 \mathscr{L}_{cos} + \beta_3 \mathscr{L}_{cls}\left(y_{rev}, \hat{y}'\right)$
  - Where $\beta_{1,2,3}$ are trade-off factors in [0,1]
- Average the loss of samples in each mini-batch before backpropagation.

Experiments

Dataset and Settings

Number of	Weibo			Twitter		
	Train	Val	Test	Train	Val	Test
Fake News	1,896	632	633	3,419	1,140	1,140
Real News	1,920	640	641	5,406	1,802	1,802
Total	3,816	1,272	1,274	8,825	2,942	2,942
		(6,362)			(14,709)	
Relevant Articles		17,849			12,419	

- Weibo-20
- 6:2:2 train : validation : test
- Relevant Articles
  - Fact-checking articles crawled from multiple fact-checking website JiaoZhen, Zhuoyaoji and Baidu Piyao.
  - Crawl other articles from Baidu News with the keywords in the Weibo posts as queries.

Experiments

Dataset and Settings

Number of	Weibo			Twitter		
	Train	Val	Test	Train	Val	Test
Fake News	1,896	632	633	3,419	1,140	1,140
Real News	1,920	640	641	5,406	1,802	1,802
Total	3,816	1,272 (6,362)	1,274	8,825	2,942 (14,709)	2,942
Relevant Articles	17,849			12,419		

- Twitter
- 6:2:2 train : validation : test
- Combine two dataset (from Snopes)
  - Utilize PHEME dataset as a supplement.
- Relevant Articles
  - Use news titles as queries and search on Google News using GNews.

Experiments

Baselines

- Pattern-based Methods: Bi-LSTM, EANN-Text,
  - BERT-Emo: uses BERT to encode the text and captures the emotion that news publisher expresses.
- Fact-based Methods:
  - DeClarE: use claim-specific attention to focus on salient words in relevant articles.
  - EVIN: evidence inference network, captures the semantic conflicts between the post and relevant articles using attention mechanism.
  - MAC: hierarchical multi-head attentive network that combines word- and article-level attention.

Experiments
Evaluation Questions

- EQ1: Can Pref-FEND improve the performance of fake news detection models with single preference?
- EQ2: Can Pref-FEND improve the performance for fake news detection that is integrated by pattern- and fact-based models?
- EQ3: How effective are the designed components of Pref-FEND?
- EQ4: How different are the Fact and the Pattern Preference Map?

Experiments

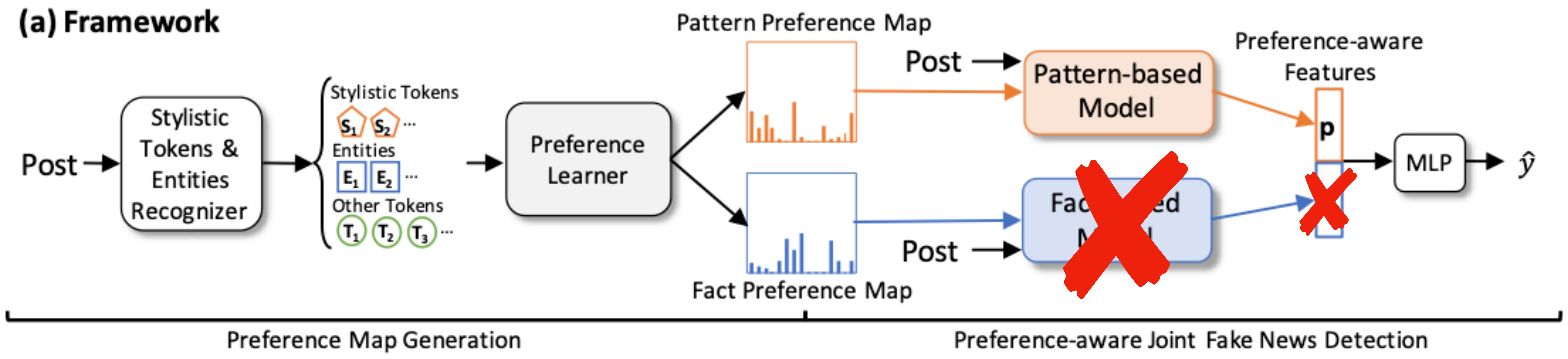
Evaluation Questions

- EQ1: Can Pref-FEND improve the performance of fake news detection models with single preference?
- EQ2: Can Pref-FEND improve the performance for fake news detection that is integrated by pattern- and fact-based models?
- EQ3: How effective are the designed components of Pref-FEND?
- EQ4: How different are the Fact and the Pattern Preference Map?

Experiments

Comparing w/ Pattern- and Fact-based Methods

- To fairly compare with existing single preference models, reduce proposed framework to a single-model version named $\mathbf{Pref-FEND}_S$.
- When comparing with a pattern-based model, remove the fact-based model but preserve the Fact Preference Map for training; and vice versa.



Experiments

Comparing w/ Pattern- and Fact-based Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Pattern-based																
Bi-LSTM	0.667	0.660	0.626	0.820	0.710	0.744	0.516	0.610	0.767	0.732	0.753	0.923	0.829	0.811	0.522	0.635
w/ Pref-FEND$_S$	**0.709**	**0.709**	0.696	0.735	0.715	0.723	0.683	0.702	**0.793**	**0.788**	0.870	0.779	0.822	0.700	0.816	0.754
EANN-Text	0.692	0.690	0.860	0.785	0.717	0.739	0.601	0.663	0.770	0.725	0.742	0.960	0.837	0.881	0.472	0.614
w/ Pref-FEND$_S$	**0.740**	**0.740**	0.760	0.697	0.727	0.723	0.783	0.752	**0.798**	**0.788**	0.837	0.832	0.834	0.737	0.744	0.741
BERT-Emo	0.712	0.708	0.667	0.839	0.743	0.787	0.587	0.672	0.794	0.762	0.769	0.950	0.850	0.873	0.550	0.675
w/ Pref-FEND$_S$	**0.746**	**0.744**	0.703	0.847	0.768	0.811	0.647	0.720	**0.804**	**0.776**	0.781	0.945	0.855	0.870	0.582	0.697
Fact-based																
DeClarE	0.684	0.678	0.642	0.820	0.720	0.755	0.549	0.636	0.786	0.753	0.765	0.941	0.844	0.853	0.543	0.663
w/ Pref-FEND$_S$	**0.706**	**0.701**	0.661	0.840	0.740	0.785	0.574	0.663	**0.798**	**0.785**	0.823	0.854	0.838	0.754	0.710	0.731
EVIN	0.707	0.706	0.683	0.768	0.690	0.738	0.647	0.690	0.783	0.761	0.788	0.884	0.833	0.773	0.623	0.690
w/ Pref-FEND$_S$	**0.712**	**0.711**	0.682	0.787	0.731	0.752	0.638	0.690	**0.795**	**0.774**	0.794	0.899	0.843	0.797	0.631	0.705
MAC	0.724	0.723	0.695	0.793	0.741	0.763	0.657	0.706	0.791	0.764	0.777	0.924	0.844	0.829	0.581	0.683
w/ Pref-FEND$_S$	**0.749**	**0.748**	0.728	0.790	0.758	0.773	0.708	0.739	**0.804**	**0.784**	0.800	0.907	0.850	0.814	0.642	0.718

- $\mathbf{Pref-FEND}_S$ successfully improves the performance of all the pattern-based and fact-based models on the two datasets.

Experiments

Comparing w/ Pattern- and Fact-based Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$\mathrm{F1}_{fake}$	P_{real}	R_{real}	$\mathrm{F1}_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$\mathrm{F1}_{fake}$	P_{real}	R_{real}	$\mathrm{F1}_{real}$
Pattern-based																
Bi-LSTM	0.667	0.660	0.626	0.820	0.710	0.744	0.516	0.610	0.767	0.732	0.753	0.923	0.829	0.811	0.522	0.635
w/ Pref-FEND$_S$	**0.709**	**0.709**	0.696	0.735	0.715	0.723	0.683	0.702	**0.793**	**0.788**	0.870	0.779	0.822	0.700	0.816	0.754
EANN-Text	0.692	0.690	0.860	0.785	0.717	0.739	0.601	0.663	0.770	0.725	0.742	0.960	0.837	0.881	0.472	0.614
w/ Pref-FEND$_S$	**0.740**	**0.740**	0.760	0.697	0.727	0.723	0.783	0.752	**0.798**	**0.788**	0.837	0.832	0.834	0.737	0.744	0.741
BERT-Emo	0.712	0.708	0.667	0.839	0.743	0.787	0.587	0.672	0.794	0.762	0.769	0.950	0.850	0.873	0.550	0.675
w/ Pref-FEND$_S$	**0.746**	**0.744**	0.703	0.847	0.768	0.811	0.647	0.720	**0.804**	**0.776**	0.781	0.945	0.855	0.870	0.582	0.697
Fact-based																
DeClarE	0.684	0.678	0.642	0.820	0.720	0.755	0.549	0.636	0.786	0.753	0.765	0.941	0.844	0.853	0.543	0.663
w/ Pref-FEND$_S$	**0.706**	**0.701**	0.661	0.840	0.740	0.785	0.574	0.663	**0.798**	**0.785**	0.823	0.854	0.838	0.754	0.710	0.731
EVIN	0.707	0.706	0.683	0.768	0.690	0.738	0.647	0.690	0.783	0.761	0.788	0.884	0.833	0.773	0.623	0.690
w/ Pref-FEND$_S$	**0.712**	**0.711**	0.682	0.787	0.731	0.752	0.638	0.690	**0.795**	**0.774**	0.794	0.899	0.843	0.797	0.631	0.705
MAC	0.724	0.723	0.695	0.793	0.741	0.763	0.657	0.706	0.791	0.764	0.777	0.924	0.844	0.829	0.581	0.683
w/ Pref-FEND$_S$	**0.749**	**0.748**	0.728	0.790	0.758	0.773	0.708	0.739	**0.804**	**0.784**	0.800	0.907	0.850	0.814	0.642	0.718

- Verifies that the original based models might be distracted from non-preferred information, thus limits their generalizability to unseen samples.

Experiments

Comparing w/ Pattern- and Fact-based Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Pattern-based																
Bi-LSTM	0.667	0.660	0.626	0.820	0.710	0.744	0.516	0.610	0.767	0.732	0.753	0.923	0.829	0.811	0.522	0.635
w/ Pref-FEND$_S$	**0.709**	**0.709**	0.696	0.735	0.715	0.723	0.683	0.702	**0.793**	**0.788**	0.870	0.779	0.822	0.700	0.816	0.754
EANN-Text	0.692	0.690	0.860	0.785	0.717	0.739	0.601	0.663	0.770	0.725	0.742	0.960	0.837	0.881	0.472	0.614
w/ Pref-FEND$_S$	**0.740**	**0.740**	0.760	0.697	0.727	0.723	0.783	0.752	**0.798**	**0.788**	0.837	0.832	0.834	0.737	0.744	0.741
BERT-Emo	0.712	0.708	0.667	0.839	0.743	0.787	0.587	0.672	0.794	0.762	0.769	0.950	0.850	0.873	0.550	0.675
w/ Pref-FEND$_S$	**0.746**	**0.744**	0.703	0.847	0.768	0.811	0.647	0.720	**0.804**	**0.776**	0.781	0.945	0.855	0.870	0.582	0.697
Fact-based																
DeClarE	0.684	0.678	0.642	0.820	0.720	0.755	0.549	0.636	0.786	0.753	0.765	0.941	0.844	0.853	0.543	0.663
w/ Pref-FEND$_S$	**0.706**	**0.701**	0.661	0.840	0.740	0.785	0.574	0.663	**0.798**	**0.785**	0.823	0.854	0.838	0.754	0.710	0.731
EVIN	0.707	0.706	0.683	0.768	0.690	0.738	0.647	0.690	0.783	0.761	0.788	0.884	0.833	0.773	0.623	0.690
w/ Pref-FEND$_S$	**0.712**	**0.711**	0.682	0.787	0.731	0.752	0.638	0.690	**0.795**	**0.774**	0.794	0.899	0.843	0.797	0.631	0.705
MAC	0.724	0.723	0.695	0.793	0.741	0.763	0.657	0.706	0.791	0.764	0.777	0.924	0.844	0.829	0.581	0.683
w/ Pref-FEND$_S$	**0.749**	**0.748**	0.728	0.790	0.758	0.773	0.708	0.739	**0.804**	**0.784**	0.800	0.907	0.850	0.814	0.642	0.718

- BERT-Emo > Bi-LTSM & EANN-Text, because BERT can generate expressive representations and the additional emotion-related features are proved helpful for this task.

Experiments

Comparing w/ Pattern- and Fact-based Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Pattern-based																
Bi-LSTM	0.667	0.660	0.626	0.820	0.710	0.744	0.516	0.610	0.767	0.732	0.753	0.923	0.829	0.811	0.522	0.635
w/ Pref-FEND$_S$	**0.709**	**0.709**	0.696	0.735	0.715	0.723	0.683	0.702	**0.793**	**0.788**	0.870	0.779	0.822	0.700	0.816	0.754
EANN-Text	0.692	0.690	0.860	0.785	0.717	0.739	0.601	0.663	0.770	0.725	0.742	0.960	0.837	0.881	0.472	0.614
w/ Pref-FEND$_S$	**0.740**	**0.740**	0.760	0.697	0.727	0.723	0.783	0.752	**0.798**	**0.788**	0.837	0.832	0.834	0.737	0.744	0.741
BERT-Emo	0.712	0.708	0.667	0.839	0.743	0.787	0.587	0.672	0.794	0.762	0.769	0.950	0.850	0.873	0.550	0.675
w/ Pref-FEND$_S$	**0.746**	**0.744**	0.703	0.847	0.768	0.811	0.647	0.720	**0.804**	**0.776**	0.781	0.945	0.855	0.870	0.582	0.697
Fact-based																
DeClarE	0.684	0.678	0.642	0.820	0.720	0.755	0.549	0.636	0.786	0.753	0.765	0.941	0.844	0.853	0.543	0.663
w/ Pref-FEND$_S$	**0.706**	**0.701**	0.661	0.840	0.740	0.785	0.574	0.663	**0.798**	**0.785**	0.823	0.854	0.838	0.754	0.710	0.731
EVIN	0.707	0.706	0.683	0.768	0.690	0.738	0.647	0.690	0.783	0.761	0.788	0.884	0.833	0.773	0.623	0.690
w/ Pref-FEND$_S$	**0.712**	**0.711**	0.682	0.787	0.731	0.752	0.638	0.690	**0.795**	**0.774**	0.794	0.899	0.843	0.797	0.631	0.705
MAC	0.724	0.723	0.695	0.793	0.741	0.763	0.657	0.706	0.791	0.764	0.777	0.924	0.844	0.829	0.581	0.683
w/ Pref-FEND$_S$	**0.749**	**0.748**	0.728	0.790	0.758	0.773	0.708	0.739	**0.804**	**0.784**	0.800	0.907	0.850	0.814	0.642	0.718

- With guidance of $\mathbf{Pref-FEND}_S$, it gains a boost. This reveals the importance of preference modeling for alleviating the overfitting of specific features.

Experiments

Comparing w/ Pattern- and Fact-based Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Pattern-based																
Bi-LSTM	0.667	0.660	0.626	0.820	0.710	0.744	0.516	0.610	0.767	0.732	0.753	0.923	0.829	0.811	0.522	0.635
w/ Pref-FEND$_S$	**0.709**	**0.709**	0.696	0.735	0.715	0.723	0.683	0.702	**0.793**	**0.788**	0.870	0.779	0.822	0.700	0.816	0.754
EANN-Text	0.692	0.690	0.860	0.785	0.717	0.739	0.601	0.663	0.770	0.725	0.742	0.960	0.837	0.881	0.472	0.614
w/ Pref-FEND$_S$	**0.740**	**0.740**	0.760	0.697	0.727	0.723	0.783	0.752	**0.798**	**0.788**	0.837	0.832	0.834	0.737	0.744	0.741
BERT-Emo	0.712	0.708	0.667	0.839	0.743	0.787	0.587	0.672	0.794	0.762	0.769	0.950	0.850	0.873	0.550	0.675
w/ Pref-FEND$_S$	**0.746**	**0.744**	0.703	0.847	0.768	0.811	0.647	0.720	**0.804**	**0.776**	0.781	0.945	0.855	0.870	0.582	0.697
Fact-based																
DeClarE	0.684	0.678	0.642	0.820	0.720	0.755	0.549	0.636	0.786	0.753	0.765	0.941	0.844	0.853	0.543	0.663
w/ Pref-FEND$_S$	**0.706**	**0.701**	0.661	0.840	0.740	0.785	0.574	0.663	**0.798**	**0.785**	0.823	0.854	0.838	0.754	0.710	0.731
EVIN	0.707	0.706	0.683	0.768	0.690	0.738	0.647	0.690	0.783	0.761	0.788	0.884	0.833	0.773	0.623	0.690
w/ Pref-FEND$_S$	**0.712**	**0.711**	0.682	0.787	0.731	0.752	0.638	0.690	**0.795**	**0.774**	0.794	0.899	0.843	0.797	0.631	0.705
MAC	0.724	0.723	0.695	0.793	0.741	0.763	0.657	0.706	0.791	0.764	0.777	0.924	0.844	0.829	0.581	0.683
w/ Pref-FEND$_S$	**0.749**	**0.748**	0.728	0.790	0.758	0.773	0.708	0.739	**0.804**	**0.784**	0.800	0.907	0.850	0.814	0.642	0.718

- MAC > DeClarE & EVIN, because it effectively uses multi-head attention to capture multi-aspect information, also can be alleviated by $\mathbf{Pref-FEND}_S$.

Experiments

Evaluation Questions

- EQ1: Can Pref-FEND improve the performance of fake news detection models with single preference?
- EQ2: Can Pref-FEND improve the performance for fake news detection that is integrated by pattern- and fact-based models?
- EQ3: How effective are the designed components of Pref-FEND?
- EQ4: How different are the Fact and the Pattern Preference Map?

Experiments

Comparing w/ Integrated (Pattern- and Fact-based) Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Last-layer Fusion	0.697	0.696	0.721	0.637	0.676	0.678	0.757	0.715	0.798	0.768	0.775	0.945	0.851	0.866	0.566	0.685
Logits Average	0.692	0.685	0.646	0.840	0.730	0.776	0.544	0.640	0.784	0.750	0.762	0.943	0.843	0.855	0.534	0.657
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
BERT-Emo (Pattern-based) + MAC (Fact-based)																
Last-layer Fusion	0.735	0.731	0.683	0.874	0.766	0.828	0.599	0.695	0.804	0.798	0.871	0.798	0.833	0.718	0.813	0.763
Logits Average	0.736	0.734	0.693	0.842	0.760	0.802	0.632	0.707	0.778	0.741	0.754	0.946	0.839	0.857	0.514	0.642
Pref-FEND	**0.756**	**0.754**	0.714	0.848	0.775	0.816	0.665	0.733	**0.814**	**0.801**	0.829	0.877	0.853	0.786	0.715	0.749

- Last-layer fusion: uses the post as input and concatenates the last-layer features of two models for final prediction.
- Logits Average: averages the models' logits (in [0,1]) for final prediction.

Experiments

Comparing w/ Integrated (Pattern- and Fact-based) Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Last-layer Fusion	0.697	0.696	0.721	0.637	0.676	0.678	0.757	0.715	0.798	0.768	0.775	0.945	0.851	0.866	0.566	0.685
Logits Average	0.692	0.685	0.646	0.840	0.730	0.776	0.544	0.640	0.784	0.750	0.762	0.943	0.843	0.855	0.534	0.657
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
BERT-Emo (Pattern-based) + MAC (Fact-based)																
Last-layer Fusion	0.735	0.731	0.683	0.874	0.766	0.828	0.599	0.695	0.804	0.798	0.871	0.798	0.833	0.718	0.813	0.763
Logits Average	0.736	0.734	0.693	0.842	0.760	0.802	0.632	0.707	0.778	0.741	0.754	0.946	0.839	0.857	0.514	0.642
Pref-FEND	**0.756**	**0.754**	0.714	0.848	0.775	0.816	0.665	0.733	**0.814**	**0.801**	0.829	0.877	0.853	0.786	0.715	0.749

- Pref-FEND >> two pattern- and fact-based methods,
  - Validates its effectiveness for integrating pattern- and fact-based models.

Experiments

Comparing w/ Integrated (Pattern- and Fact-based) Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Last-layer Fusion	0.697	0.696	0.721	0.637	0.676	0.678	0.757	0.715	0.798	0.768	0.775	0.945	0.851	0.866	0.566	0.685
Logits Average	0.692	0.685	0.646	0.840	0.730	0.776	0.544	0.640	0.784	0.750	0.762	0.943	0.843	0.855	0.534	0.657
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
Bi-LSTM	0.667	0.660	0.626	0.820	0.710	0.744	0.516	0.610	0.767	0.732	0.753	0.923	0.829	0.811	0.522	0.635
w/ Pref-FEND$_S$	**0.709**	**0.709**	0.696	0.735	0.715	0.723	0.683	0.702	**0.793**	**0.788**	0.870	0.779	0.822	0.700	0.816	0.754
DeClarE	0.684	0.678	0.642	0.820	0.720	0.755	0.549	0.636	0.786	0.753	0.765	0.941	0.844	0.853	0.543	0.663
w/ Pref-FEND$_S$	**0.706**	**0.701**	0.661	0.840	0.740	0.785	0.574	0.663	**0.798**	**0.785**	0.823	0.854	0.838	0.754	0.710	0.731

- Pref-FEND brings further improvements based on the remarkable performance of $\mathbf{Pref-FEND}_S$ w.r.t. the same base models.
- This proves that Pref-FEND is applicable to both the single-preference models and the integrated models based on them.

Experiments

Comparing w/ Integrated (Pattern- and Fact-based) Methods

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Last-layer Fusion	0.697	0.696	0.721	0.637	0.676	0.678	0.757	0.715	0.798	0.768	0.775	0.945	0.851	0.866	0.566	0.685
Logits Average	0.692	0.685	0.646	0.840	0.730	0.776	0.544	0.640	0.784	0.750	0.762	0.943	0.843	0.855	0.534	0.657
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
BERT-Emo (Pattern-based) + MAC (Fact-based)																
Last-layer Fusion	0.735	0.731	0.683	0.874	0.766	0.828	0.599	0.695	0.804	0.798	0.871	0.798	0.833	0.718	0.813	0.763
Logits Average	0.736	0.734	0.693	0.842	0.760	0.802	0.632	0.707	0.778	0.741	0.754	0.946	0.839	0.857	0.514	0.642
Pref-FEND	**0.756**	**0.754**	0.714	0.848	0.775	0.816	0.665	0.733	**0.814**	**0.801**	0.829	0.877	0.853	0.786	0.715	0.749

- Last-layer fusion does not necessarily perform better than the simple logits average.
  - Indicates that last-layer fusion may be insufficient to align the feature spaces of the pattern- and the fact-based model, which leads to negative fusion effects.

Experiments
Evaluation Questions

- EQ1: Can Pref-FEND improve the performance of fake news detection models with single preference?
- EQ2: Can Pref-FEND improve the performance for fake news detection that is integrated by pattern- and fact-based models?
- EQ3: How effective are the designed components of Pref-FEND?
- EQ4: How different are the Fact and the Pattern Preference Map?

Experiments

Effectiveness of Model Preference Learning

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
w/ rand init maps	0.694	0.693	0.676	0.736	0.705	0.715	0.652	0.682	0.788	0.765	0.787	0.896	0.838	0.790	0.616	0.692
w/o $\mathcal{L}_{cos}$	0.701	0.703	0.672	0.787	0.725	0.747	0.621	0.678	0.794	0.785	0.845	0.813	0.829	0.721	0.764	0.742
w/o $\mathcal{L}_{cls}(y_{rev}, \hat{y}')$	0.703	0.702	0.710	0.679	0.694	0.696	0.725	0.710	0.792	0.764	0.775	0.932	0.846	0.842	0.571	0.681
w/ only $\mathcal{L}_{cls}(y, \hat{y})$	0.700	0.702	0.672	0.782	0.723	0.743	0.622	0.677	0.789	0.747	0.752	0.979	0.851	0.936	0.490	0.643
BERT-Emo (Pattern-based) + MAC (Fact-based)																
Pref-FEND	**0.756**	**0.754**	0.714	0.848	0.775	0.816	0.665	0.733	**0.814**	**0.801**	0.829	0.877	0.853	0.786	0.715	0.749
w/ rand init maps	0.723	0.716	0.666	0.886	0.761	0.833	0.562	0.671	0.806	0.786	0.801	0.911	0.852	0.820	0.642	0.720
w/o $\mathcal{L}_{cos}$	0.747	0.745	0.706	0.842	0.768	0.807	0.654	0.722	0.807	0.801	0.874	0.799	0.835	0.721	0.819	0.767
w/o $\mathcal{L}_{cls}(y_{rev}, \hat{y}')$	0.745	0.740	0.690	0.883	0.775	0.841	0.608	0.706	0.808	0.789	0.806	0.903	0.852	0.811	0.657	0.726
w/ only $\mathcal{L}_{cls}(y, \hat{y})$	0.741	0.735	0.682	0.896	0.775	0.851	0.588	0.696	0.792	0.787	0.869	0.778	0.821	0.699	0.815	0.752

- Randomly initialize preference maps, forces the generation of preference maps to rely on the supervision of ground-truth labels.

Experiments

Effectiveness of Model Preference Learning

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
w/ rand init maps	0.694	0.693	0.676	0.736	0.705	0.715	0.652	0.682	0.788	0.765	0.787	0.896	0.838	0.790	0.616	0.692
w/o $\mathcal{L}_{cos}$	0.701	0.703	0.672	0.787	0.725	0.747	0.621	0.678	0.794	0.785	0.845	0.813	0.829	0.721	0.764	0.742
w/o $\mathcal{L}_{cls}(y_{rev}, \hat{y}\prime)$	0.703	0.702	0.710	0.679	0.694	0.696	0.725	0.710	0.792	0.764	0.775	0.932	0.846	0.842	0.571	0.681
w/ only $\mathcal{L}_{cls}(y, \hat{y})$	0.700	0.702	0.672	0.782	0.723	0.743	0.622	0.677	0.789	0.747	0.752	0.979	0.851	0.936	0.490	0.643
BERT-Emo (Pattern-based) + MAC (Fact-based)																
Pref-FEND	**0.756**	**0.754**	0.714	0.848	0.775	0.816	0.665	0.733	**0.814**	**0.801**	0.829	0.877	0.853	0.786	0.715	0.749
w/ rand init maps	0.723	0.716	0.666	0.886	0.761	0.833	0.562	0.671	0.806	0.786	0.801	0.911	0.852	0.820	0.642	0.720
w/o $\mathcal{L}_{cos}$	0.747	0.745	0.706	0.842	0.768	0.807	0.654	0.722	0.807	0.801	0.874	0.799	0.835	0.721	0.819	0.767
w/o $\mathcal{L}_{cls}(y_{rev}, \hat{y}\prime)$	0.745	0.740	0.690	0.883	0.775	0.841	0.608	0.706	0.808	0.789	0.806	0.903	0.852	0.811	0.657	0.726
w/ only $\mathcal{L}_{cls}(y, \hat{y})$	0.741	0.735	0.682	0.896	0.775	0.851	0.588	0.696	0.792	0.787	0.869	0.778	0.821	0.699	0.815	0.752

- It falls behind the complete Pref-FEND, proves the effectiveness of model preference learning, which exploits prior knowledge in a dynamic graph representation learning process.

Experiments

Effectiveness of Losses for Differentiating the Preference Maps

Method	Weibo								Twitter							
	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$	Acc.	macF1	P_{fake}	R_{fake}	$F1_{fake}$	P_{real}	R_{real}	$F1_{real}$
Bi-LSTM (Pattern-based) + DeClarE (Fact-based)																
Pref-FEND	**0.714**	**0.712**	0.684	0.788	0.732	0.754	0.640	0.692	**0.812**	**0.792**	0.803	0.917	0.857	0.832	0.645	0.727
w/ rand init maps	0.694	0.693	0.676	0.736	0.705	0.715	0.652	0.682	0.788	0.765	0.787	0.896	0.838	0.790	0.616	0.692
w/o $\mathcal{L}_{cos}$	0.701	0.703	0.672	0.787	0.725	0.747	0.621	0.678	0.794	0.785	0.845	0.813	0.829	0.721	0.764	0.742
w/o $\mathcal{L}_{cls}(y_{rev}, \hat{y}\prime)$	0.703	0.702	0.710	0.679	0.694	0.696	0.725	0.710	0.792	0.764	0.775	0.932	0.846	0.842	0.571	0.681
w/ only $\mathcal{L}_{cls}(y, \hat{y})$	0.700	0.702	0.672	0.782	0.723	0.743	0.622	0.677	0.789	0.747	0.752	0.979	0.851	0.936	0.490	0.643
BERT-Emo (Pattern-based) + MAC (Fact-based)																
Pref-FEND	**0.756**	**0.754**	0.714	0.848	0.775	0.816	0.665	0.733	**0.814**	**0.801**	0.829	0.877	0.853	0.786	0.715	0.749
w/ rand init maps	0.723	0.716	0.666	0.886	0.761	0.833	0.562	0.671	0.806	0.786	0.801	0.911	0.852	0.820	0.642	0.720
w/o $\mathcal{L}_{cos}$	0.747	0.745	0.706	0.842	0.768	0.807	0.654	0.722	0.807	0.801	0.874	0.799	0.835	0.721	0.819	0.767
w/o $\mathcal{L}_{cls}(y_{rev}, \hat{y}\prime)$	0.745	0.740	0.690	0.883	0.775	0.841	0.608	0.706	0.808	0.789	0.806	0.903	0.852	0.811	0.657	0.726
w/ only $\mathcal{L}_{cls}(y, \hat{y})$	0.741	0.735	0.682	0.896	0.775	0.851	0.588	0.696	0.792	0.787	0.869	0.778	0.821	0.699	0.815	0.752

- The largest drop occur when removing both the two losses, indicates that the auxiliary losses are effective and necessary to generate better preference map for integration of models with different preferences.

Experiments
Evaluation Questions

- EQ1: Can Pref-FEND improve the performance of fake news detection models with single preference?
- EQ2: Can Pref-FEND improve the performance for fake news detection that is integrated by pattern- and fact-based models?
- EQ3: How effective are the designed components of Pref-FEND?
- EQ4: How different are the Fact and the Pattern Preference Map?

Experiments

Case study

Pattern-preferred token

Fact-preferred token

#	Post (Translated into English)
1	A group of city administration officials in Sishui , Shandong , chased an old man until all his eggs were broken on the ground . The old man sat there helplessly . The officials ran away after hitting . The white-haired man should be about 80 years old , and he can't make much money by selling eggs . So why be aggressive ? Is there no moment for the officials to be alone ? If the officials only oppresses citizens , what's the good of having these officials ? You will be punished sooner or later for bullying the underprivileged .
	Ground Truth: Fake **Judgment:** Bi-LSTM (Fake), DeClarE (Real), Pref-FEND (Fake)
2	[A student of ZJU jumping to the West Lake for a crazy graduation photo drowned] On June 29 , Xin (not his real name) from ZJU and his classmates went to the waters near the scenic spot of " Konggu Chuanyin " in Gushan , Beili Lake , West Lake in Hangzhou . Xin asked his classmates to take pictures of his swimming underwater . He jumped into the West Lake from the side of Xiling brige on Beishan Road and swam to the lotus pool of Gushan park on the other side . He drowned when swimming to the center of the lake . Recently , he has received a full PhD scholarship from a U.S. university .
	Ground Truth: Fake **Judgment:** Bi-LSTM (Real), DeClarE (Fake), Pref-FEND (Fake)
3	Is anyone in Shanghai interested in raising a dog ? No Charge . Golden Retriever , Poodle , Samoyed , and other more breeds . There are dog-killing slaughterhouses being destroyed . If no one adopts , they will be euthanized . Let these little cute lives accompany with you . If you are really not able to raise them , please forward this message .
	Ground Truth: Fake **Judgment:** Bi-LSTM (Real), DeClarE (Real), Pref-FEND (Fake)

- Case 1 conveys strong signals of emotional patterns, which are preferred by pattern-based models.
- Case 2 contains a large number of places and event descriptions, which is friendly to utilize the evidential texts in relevant articles.
- Due to the different dominant signals, the pattern-based Bi-LSTM judges correctly in Case 1, but fails in Case 2, and the judgments of the fact-based DeClarE are the opposite.

Experiments

Case study

#	Post (Translated into English)
1	A group of city administration officials in Sishui , Shandong , chased an old man until all his eggs were broken on the ground . The old man sat there helplessly . The officials ran away after hitting . The white-haired man should be about 80 years old , and he can't make much money by selling eggs . So why be aggressive ? Is there no moment for the officials to be alone ? If the officials only oppresses citizens , what's the good of having these officials ? You will be punished sooner or later for bullying the underprivileged .
	Ground Truth: Fake **Judgment:** Bi-LSTM (Fake), DeClarE (Real), Pref-FEND (Fake)
2	[A student of ZJU jumping to the West Lake for a crazy graduation photo drowned] On June 29 , Xin (not his real name) from ZJU and his classmates went to the waters near the scenic spot of " Konggu Chuanyin " in Gushan , Beili Lake , West Lake in Hangzhou . Xin asked his classmates to take pictures of his swimming underwater . He jumped into the West Lake from the side of Xiling brige on Beishan Road and swam to the lotus pool of Gushan park on the other side . He drowned when swimming to the center of the lake . Recently , he has received a full PhD scholarship from a U.S. university .
	Ground Truth: Fake **Judgment:** Bi-LSTM (Real), DeClarE (Fake), Pref-FEND (Fake)
3	Is anyone in Shanghai interested in raising a dog ? No Charge . Golden Retriever , Poodle , Samoyed , and other more breeds . There are dog-killing slaughterhouses being destroyed . If no one adopts , they will be euthanized . Let these little cute lives accompany with you . If you are really not able to raise them , please forward this message .
	Ground Truth: Fake **Judgment:** Bi-LSTM (Real), DeClarE (Real), Pref-FEND (Fake)

Pattern-preferred token

Fact-preferred token

- In Case 3, both of them wrongly judge this post as real.
- Speculate that the failure is led by the negative inference from the non-preferred information.
- With help of model preference learning, Pref-FEND succeed in judging all three posts as fake.

Conclusions and Future Work

- Propose Pref-FEND to integrate the pattern- & fact-based FND models in a preference-aware fashion.
    - The learned preference maps guide the models to focus more on their preferred parts with less interference by non-preferred parts.
- How to enhance the interaction between the preference map generation and specific models and how to extend the framework to multi-class and multi-preference scenarios are expected.

Comments

of Pref-FEND

- Also Text-only method.
- Focus on integration of pattern- & fact-based detection model.
- In case study, select single preference method that performance lowest to compare.
    - Not fair, maybe BERT-Emo / MAC can recognize fake news correctly.
- Concept of modeling preference map is good.
    - Also consider to differ two map in design loss function.