# Social-PTM: Pre-training Model for Social Media Language Understanding using Hierarchical Contextual Constructive Learning Framework

**Anonymous submission**

**Abstract**

Large-scale pre-training models (PTMs) are getting increasingly success. However, the domain gap between the social media language and the formal language hinders these PTMs from being well-utilized in Social Media Language Understanding. In this paper, we attribute the domain gap to two aspects: in terms of Vocabulary and Grammar Use, abbreviations, emojis and even spelling mistakes are common in social media language, while they are rarely in formal language; in terms of Context Information Distribution, unlike the formal language which enjoys rich and continuous context information, the sparse context information distribution in social media language makes it hard for using present pseudo tasks to train PTMs. So we model the social media language with the tree structure which can facilitate in context information modeling, view a post and its comments as a single unit, where the post serves as the root node, the comments serve as the child nodes and the reply operations serve as the edges.

Based on these, We present Social-PTM, a pre-training model for Social Media Language Understanding using the Hierarchical Contextual Constructive Learning Framework (HCCLF). The HCCLF is hierarchically designed with three tasks at three levels respectively, the Mask Language Prediction (MLP) task at the word level, the Predecessor Prediction (PP) task at the intra-comments level, and the Post-Comments Semantic Constructive Learning (PCS-CL) task at the inter-post-comments level. The first task enables word-level representation learning, the other two constructive learning tasks enable context-level representation learning of social media language.

What is more, considering of the various downstream tasks in Social Media Language Understanding, we propose to categorize them from three perspectives and design a universal performance evaluation benchmark, called Social-UPEB. We expect that Social-PTM will present state-of-the-art (SOTA) performance on Social-UPEB.

**Keywords:** Natural Language Processing, Language Model, Pre-training Model, Social Media Language Understanding

## 1. Introduction

Pre-training models (PTMs) have gained great success in Natural Language Processing (NLP) since they can learn universal representations and can be easily adapted to downstream NLP tasks, especially those with scarce data (Qiu et al., 2020; Han et al., 2021). Nowadays, social media is playing a more and more crucial role in communication (Ding et al., 2020). Present mainstream PTMs, such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020) and their variants (Poerner et al., 2020; Raffel et al., 2020a), which have achieved state-of-the-art (SOTA) performance on various NLP tasks, can not be well-applied in Social Media Language Understanding. This is caused by the **domain gap between social media language and the formal language**: the corpus used to train the mainstream PTMs are formal language [1] (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020b; Touvron et al., 2023), such as books and Wikipedia pages; while social media language is not formal, it is user-generated, is sometimes short in length, and involves social engagements (Nguyen et al., 2020a; DeLucia et al., 2022; Zhang et al., 2022b; Barbieri et al., 2022).

We attribute the domain gap between the formal language and the social media language to two aspects, an example of the comparison is shown in Figure 1.

- **Vocabulary and Grammar Use.** Grammar mistakes, spelling mistakes, and inappropriate word use happen much more frequently in social media language, while grammar and vocabulary are always correct and regular in formal language (Nguyen et al., 2020b). Besides the mistakes, hashtags, abbreviations, and emojis prevail in social media language, while these are rarely seen in formal language (Nguyen et al., 2020b).
- **Context Information Distribution.** The formal language enjoys very rich and continuous context information. For one thing, inner a sentence, there is usually dozens of words; For

[1] Part of the training data of LLaMA (Touvron et al., 2023) is English CommonCrawl (Wenzek et al., 2020) and C4 (Raffel et al., 2020b), which contains social media language data, such as data from Twitter. However, LLaMA uses a linear model to filter out data that is not similar to Wikipedia data. In other words, truly, it takes into consideration of some social media language data, but it only uses the part which is similar to the formal language.

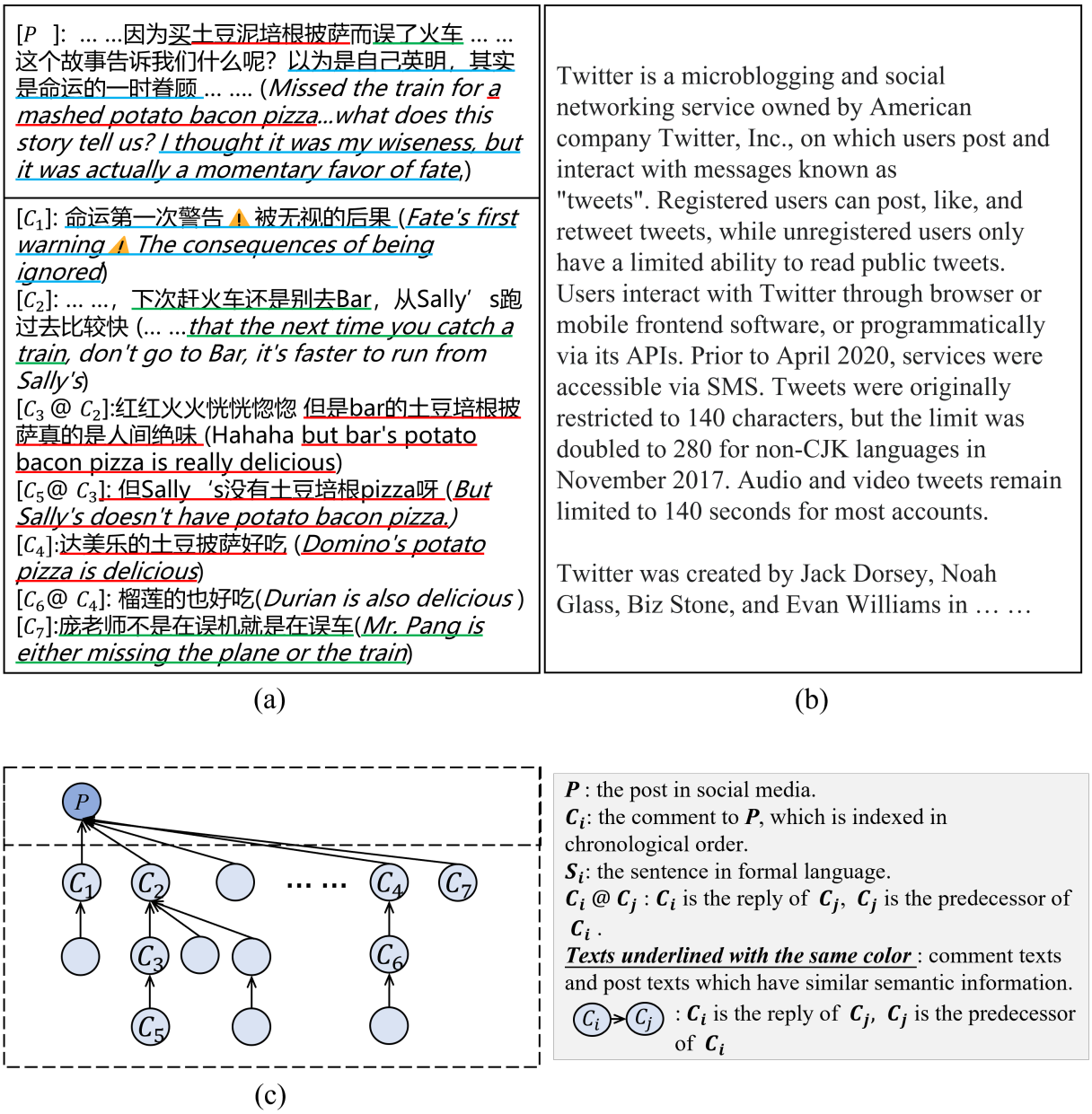


Figure 1: The domain gap between the social media language and the formal language, which is used to train mainstream PTMs. (a). Example single unit in social media language, a post-comments pair (a post and corresponding comments) from Chinese Sina Weibo, with the English translation in parentheses. (b). Example formal language text from Wikipedia. (c). Modeling Social media language with the tree structure.

In terms of Grammar and Language Use, the Emoji in $C_1$ and "红红火火恍恍惚惚" in $C_3$ show the social media language's informal characteristics. In terms of Context Structure, a post-comments pair in (a) is of the tree structure of (c), and the interaction relationship among the post and its comments is hierarchical: At the intra-comments level, among $C_1$-$C_7$, $C_6$ is the reply of $C_4$ and they both talk about pizza's flavor; At the inter-post-comments level, $P$'s semantic information can be divided into three parts, $C_1$-$C_5$ and $C_7$ each corresponds to one semantic part and they together correspond to the total semantic information of $P$.

another thing, among sentences/paragraphs, a sentence/paragraph usually has strong relationship with its prior and its latter one. This makes the mask language prediction (MLP) (Devlin et al., 2019), next sentence prediction (NSP) (Devlin et al., 2019), next token prediction (NTP) (Brown et al., 2020), Language Permutation (LP) (Yang et al., 2019), etc, suitable for language modeling. However, the context information distribution in social media language is very sparse. For one thing, for a single post/comment, it is usually very short in length, such as a dozen of words or even less then ten words. Therefore, the context information in a single post/comment is rare, and this makes it hard to build PTMs with MLP and NTP. For another thing, among the posts and comments, since there is the reply operation, a post/comment can have strong relationship with many comments. In other words, the context information is not merely constrained within a single post/comment, but scatters among the posts and comments, which might be very far away from each other if the posts/comments are just logged chronologically. Therefore, this makes it hard to build PTMs with PP and NSP.

For problem caused by the domain gap in Vocabulary and Grammar Use, it can be easily solved by using large scale domain specific data. For problem caused by the domain gap in Context Information Distribution, we tend to model the social media language with the tree structure (Benamara et al., 2018; Li et al., 2018; Zubiaga et al., 2016; Louis and Cohen, 2015; Ma et al., 2018). In this way, we can take into consideration the reply relationship, use it as a bridge to union the context information scattered among the posts/comments together, and then build the PTM for better social media language understanding.

Our tree structure modeling of social media language is shown in Figure 1(a) and (c). As for the structure, shown in Figure 1(c), we view a post and its comments as a single unit, the post serves as the root node, its comments serve as the child nodes and the reply operations serve as the edges. As for the context information, shown in Figure 1(a), at the **intra-comments level** (among a post's comments), the only interaction exists in the comment-to-comment reply relationship. At the **inter-post-comments level** (between a post and its comments set), since all comments are direct/indirect replies to the post, the post somewhat serves as the topic sentence(s) to its comments set, with all comments' total semantic information corresponding to that of the post and most single comments partially corresponding to the post.

Considering these domain-specific aspects in social media language and the tree structure modeling of social media language, we present **Social-PTM**, a pre-trained model using the proposed **Hierarchical Contextual Constructive Learning Framework (HCCLF)** for Social Media Language Understanding. To model the word-level representation, we apply the **Masked Language Prediction (MLP)** (Devlin et al., 2019). To model the context-level representation, based on social media's tree structure, we design two constructive learning (Arora et al., 2019) tasks, a **Predecessor Prediction (PP)** task for intra-comments level context modeling, and a **Post-Comments Semantic Constructive Learning (PCS-CL)** task for inter-post-comments level context modeling. The PP task is designed to reconstruct the social media language's tree structure. The PCS-CL task assumes that representations for a post-comments pair (a post and its corresponding comments set) should be similar, while those for non-pairs should be dissimilar. What's more, to model social media language's characteristics in grammar and language use, the corpus used to train Social-PTM is explicitly collected from well-known social media platforms, such as Twitter and Sina Weibo.

Beside HCCLF designing, to evaluate systematically, we propose to develop a Universal Performance Evaluation Benchmark for Social Media Language Understanding, called **Social-UPEB**. We categorize Social Media Language Understanding tasks from three perspectives: (a). token-level task/context-level task, (b). prediction task/generation task, and (c). adequate-resource task/low-resource task/zero-resource task. To achieve universal performance evaluation, the Social-UPEB is designed to cover tasks in all above mentioned areas.

## 2. Releated Work

**Social Media Language Pre-training Model.** Recent studies have been aware of the domain gap between social media language and the formal language, and they have argued the necessity of PTMs for Social Media Language Understanding (Nguyen et al., 2020a; DeLucia et al., 2022; Zhang et al., 2022b; Barbieri et al., 2022; Loureiro et al., 2022). However, they either fail to deal with the domain gap on context structure or even fail to concentrate on it.

BERTweet (Nguyen et al., 2020a), XLM-T (Barbieri et al., 2022) and Bernice(DeLucia et al., 2022) concentrate on the grammar and language use aspect. BERTweet (Nguyen et al., 2020a) concentrates on the special tokens, such as emojis and hashtags ,XLM-T (Barbieri et al., 2022) finds the multilingual characteristic of social media language, Bernice(DeLucia et al., 2022) designs a

special tokenizer, and they then just train the PTMs with domain-specific corpora. TimeLMs (Loureiro et al., 2022) asserts the importance of time variable, trains a series of PTMs in chronological order, and verifies this strategy's capacity of dealing with the future and out-of-distribution posts. TwHIN-BERT (Zhang et al., 2022b) is aware of the context structure characteristic of social media language, such as that tweet text is usually short without that much context information. Instead of exploring the nature of social media language's context structure, it uses the social engagements, such as faves, retweets, and replies, to help model the contextual information. In terms of practical method, it uses the social engagements to construct a graph, mines socially similar pairs by another algorithm, and uses constructive learning method based on these pairs to model context embedding. Since the medium for context modeling, the similar pairs, are learnt, rather than natural, we cast doubt on the reliability of the method.

**Evaluation Benchmark for Social Media Language Understanding.** Several studies (Nguyen et al., 2020a; DeLucia et al., 2022; Barbieri et al., 2022; Loureiro et al., 2022; Barbieri et al., 2020) on PTMs for Social Media Language Understanding curate their own evaluation benchmarks. These benchmarks concentrates on different aspects of social media language. We will list them, describe them, and analyze them slightly.

TweetEval (Barbieri et al., 2020) is the first evaluation benchmark for Social Media Language Understanding, and has been used by (DeLucia et al., 2022; Zhang et al., 2022b; Barbieri et al., 2022). It is made up of seven heterogeneous Tweet classification tasks: Emoji Prediction, Emotion Recognition, Hate Speech Detection, Irony Detection, Offensive Language Identification, Sentiment Analysis, and Stance Detection. These tasks are all semantic-level classification tasks.

Unified Multilingual Sentiment Analysis Benchmark (UMSAB) (Barbieri et al., 2022) mainly aims to evaluate PTMs' multilingual modeling ability, and has been used by (DeLucia et al., 2022; Zhang et al., 2022b; Barbieri et al., 2022). It is a Sentiment Analysis dataset with corpora from eight topologically distant languages. What is more, this dataset can evaluate the PTMs in two scenarios. One is the very general scenario: the PTM is fine-tuned on the train set and evaluated on the test set; The other is the zero-shot scenario: the PTM is fine-tuned on the train set of language A but evaluated on the test set of language B. In the latter scenario, the PTM can be evaluated whether and to which extend it has disentangled semantic information from the language type.

Besides, several other benchmark tasks have to be talked about. TwHIN-BERT (Zhang et al., 2022b) also collects several other semantic-level Tweet classification tasks. They are Hashtag Prediction, Topic Classification, and Social Engagement Prediction. BERTweet (Nguyen et al., 2020a) also collects two token-level tasks, the Part-of-speech (POS) tagging task, and the Named-entity Recognition task.

Last, there are various downstream tasks in Social Media Language Understanding, such as User Profiling (Heidari et al., 2020), and Empathetic Dialogue Generation (Majumder et al., 2020). The above-mentioned benchmarks only contain a minority of them. We think they should be all included, we will talk about this in Section 4.

## 3. Proposed Method

The overall architecture of the Hierarchical Contextual Constructive Learning Framework (HCCLF) for training social-PTM is shown in Figure 2(a). The HCCLF is hierarchically designed with three tasks at three levels respectively, the Mask Language Prediction (MLP) task at the word level, the Predecessor Prediction (PP) task at the intra-comments level, and the Post-Comments semantic constructive learning (PCS-CL) task at the inter-post-comments level. The first task is used for word-level representation learning, the other two are used for context-level representation learning.

Input for HCCLF is a batch of post-comments pairs. We denote $N$ as the batch size, $P^i$ as the i-th post, $C^i$ as the comments set of $P^i$. Inside $C^i$, comments are indexed chronologically as $C^i =< C^i_1, C^i_2, ....., C^i_{m^i} >, m^i = |C^i|$. Outputs are mainly post-comments pairs' inner relationships, we will talk about them in HCCLF's three tasks.

### 3.1. Social-PTM

The architecture of social-PTM is the same as BERT (Devlin et al., 2019). As seen in the left part of Figure 2(b), Social-PTM is a multi-layer bidirectional Transformer encoder (Vaswani et al., 2017). The Transformer encoder layer consists of a Multi-Head Attention sub-layer and a position-wise fully connected sub-layer, with the residual connection (He et al., 2016) and layer normalization (Ba et al., 2016) after each sub-layer. In our work, we denote the number of encoder layers as $L$, the hidden size as $H$, and attention heads num as $A$.

Input for Social-PTM is a single post/comment. Before feeding the input to Social-PTM, a zero position $< E >$ is first added at the beginning, each element in the input sequence is then tokenized and finally summed up with the position embedding, as described in (Vaswani et al., 2017). Outputs are a sequence of representations, the one corresponding to $< E >$ in position is the

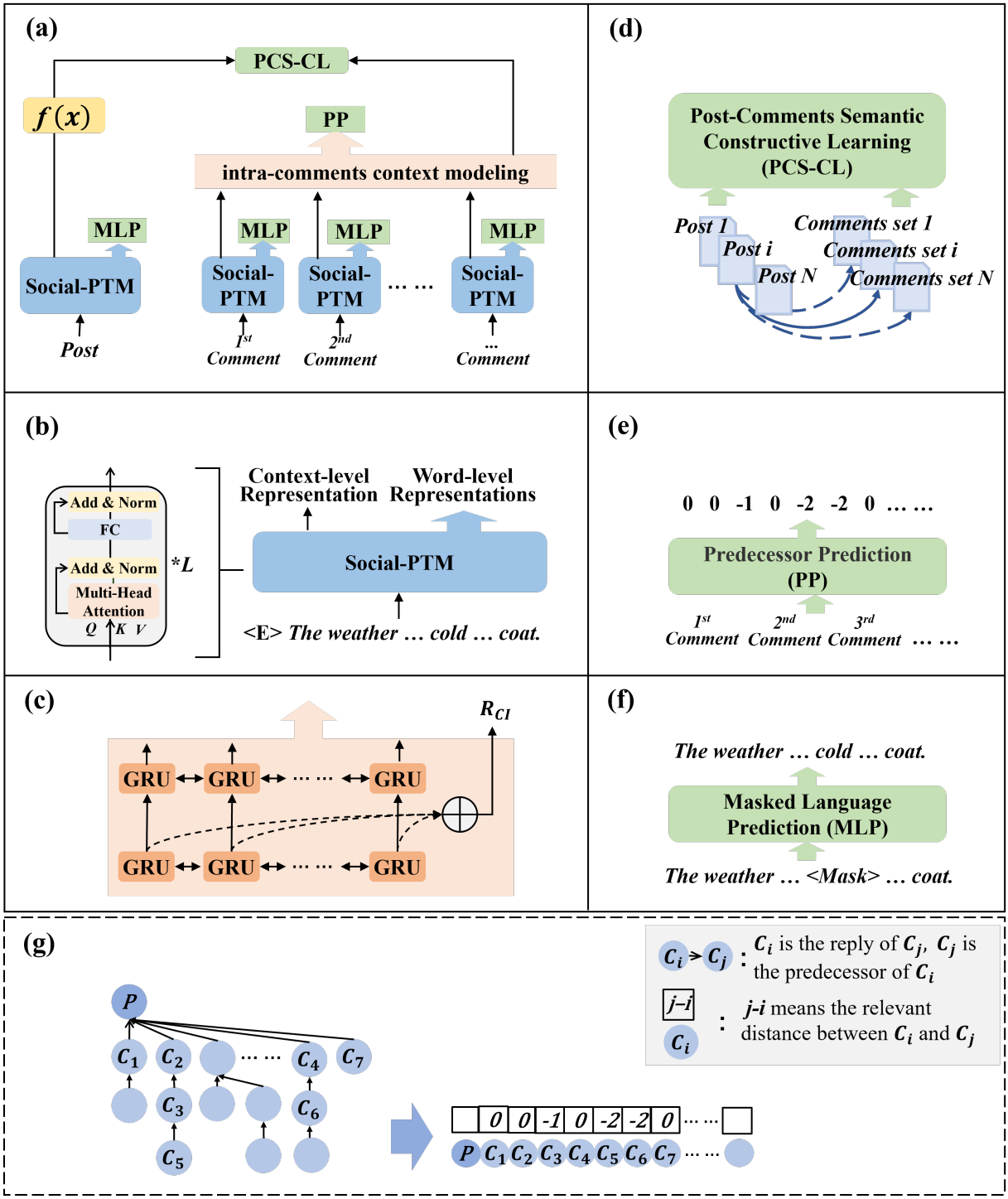


Figure 2: The proposed method. (a). Hierarchical Contextual Constructive Learning Framework (HCCLF). (b). The Social-PTM. (c). The auxiliary Intra-comments Context Modeling block (ICCMB). (d). The Post-Comments Semantic Constructive Learning (PCS-CL) task. (e). The Predecessor Prediction (PP) task. (f). The Mask Language Prediction (MLP) task. (g). The linear representation for the tree structure of a post-comments pair.

context-level representation, and the others are word-level representations with each positionally corresponding to that in the input sequence. We denote $R_P^i$ as context-level representation for the post $P^i$, $R_C^i =< r_1^i, r_2^i, ....., r_{m^i}^i >$ for comments set $C^i$.

### 3.2. Word-level Representation Learning

The Masked Language Prediction (MLP) task, as described in (Devlin et al., 2019), helps word-level representation learning by probabilistic language modeling. As seen in Figure 2(f), MLP randomly masks out some tokens in the input sequence and trains the model to predict the masked by its context.

Input for MLP is the masked word sequence $W^{mask} =< w_1, w_2, ..., < Mask >, ..., w_n >$ of a single post/comment where $n$ is the sequence length. Output is the original word sequence $W =< w_1, w_2, ..., w_i, ..., w_n >$. In our work, the masked word prediction job from $W^{mask}$ to $W$ is done by Social-PTM.

### 3.3. Context-level Representation Learning

#### 3.3.1. the auxiliary Intra-comments Context Modeling block

In HCCLF, we use the interaction relationship among posts and comments to facilitate the Social-PTM model the context-level representations. However, our Social-PTM takes exactly one single post/comment as input. To deal with this issue, we introduce an auxiliary block which will be through away right after the pre-training phase.

The architecture of the auxiliary block, the Intra-comments Context Modeling block(ICCMB), is shown in Figure 2(c). To facilitate PP, ICCMB uses the bi-directional GRU (BiGRU) (Chung et al., 2014) layers to model the intra-comments relationship. It has to mention that the layer num $N^G$ of BiGRUs should be small due to ICCMB's auxiliary nature. To facilitate PCS-CL, ICCMB uses an attention pooling operation to integrate the representations of all comments.

Inputs for ICCMB are representations of certain comments set $C$, $R_C =< r_1, r_2, ....., r_m >$, which are the direct output of the Social-PTM (for simplicity, we dismiss the superscription of the comments set index in this subsection). We denote $h^i =< h_1^i, h_2^i, ....., h_m^i > (i = 1, 2, .....N^G)$ as hidden states of the i-th BiGRU layer. Outputs consist of two parts. To facilitate PP, the outputs of intra-comments relationship are $h^{N^G} =< h_1^{N^G}, h_2^{N^G}, ....., h_m^{N^G} >$, because they are deeper in ICCMB and better for the tree structure reconstruction task of PP. To facilitate PCS-CL, the output of integrating all representations in a comments set is $R_{CI}$. ICCMB uses the first layer of hidden states $h^1$ as the source of stance output since they are closer to the original representations set $R_C$, and uses an attention pooling layer to capture contribution imbalance. $R_{CI}$ is calculated as

$$\begin{aligned} p_i &= tanh(W_p h_i^1 + b_p) \\ u_i &= \frac{e^{w_u p_i}}{\sum_j e^{w_u p_j}} \\ R_{CI} &= \sum_i u_i h_i^1, \end{aligned} \tag{1}$$

where $W_p, b_p, w_u$ are learnable parameters.

#### 3.3.2. Predecessor Prediction

The Predecessor Prediction (PP) task helps context-level representation learning by modeling the intra-comments relationship. As seen in Figure 2(e), given that all participants are known as comments, the PP task is to reconstruct the tree structure of the corresponding post-comments pair. In plain words, it is to predict the "who reply who" relationship.

An additional space with linear complexity is used to record the tree structure of a post-comments pair. In notation, given certain comments set $C$, the tree structure is recorded as $\hat{T_C} =< \hat{d_1}, \hat{d_2}, ....., \hat{d_m} >, m = |C|$ (for simplicity, we dismiss the superscription of comments set index in this subsection). As seen in Figure 2(g), in the tree structure of a post-comments pair, if $C_i$ is the reply for $C_j$, we say $C_j$ is the predecessor of $C_i$ and $C_i$ is the successor of $C_j$; In the linear representation of the tree structure, the post and corresponding comments are stored and indexed chronologically, for $C_i$, we use the relevant distance $\hat{d_i} = j - i$ to denote its predecessor's location in the comments set. What is more, when a certain comment is a direct reply to the post with no comment predecessor, we use $0$ to denote.

Inputs for PP are representations of every single comment in certain comments set $C$, $R_C =< r_1, r_2, ....., r_m >$. Output for PP is the predicted tree structure $T_C =< d_1, d_2, ....., d_m >$. In our work, the projection from $R_C$ to $T_C$ is done by ICCMB where $T_C$ in PP is actually $h^{N^G}$.

#### 3.3.3. Post-Comments Semantic Constructive Learning

The Post-Comments Semantic Constructive Learning (PCS-CL) task helps context-level representation learning, especially that of the post, by modeling the inter-post-comments relationship. As seen in Figure 2(d), with the solid line demonstrating greater correlation than the dashed line, PCS-CL constrains representations for a post-comments pair to be similar while those for non-pairs to be dissimilar.

Inputs for PCS-CL are representations of a batch of post-comments pairs, $R_P^i$ and $R_C^i =< r_1^i, r_2^i, ....., r_{m^i}^i >$ with $i \in [0, N)$, which are direct outputs of Social-PTM. There is no output in PCS-CL. As for the loss function, we design

Table 1: Description of the proposed Social-UPEB. 'Gran' means whether the task is in token-level or context-level. 'P/G' means whether the task is a prediction task (P) or a generation task (G). 'Rsc' means whether the task is an adequate-resource task, a low-resource task, or a zero-resource task, and 'any' means that the task can be any of the above three kinds. 'Example' means the specific example dataset. What is more, we do not give descriptions to some tasks, because they are easy to be understood by their name.

| Task | Description | Gran | P/G | Rsc | Example |
|---|---|---|---|---|---|
| Named-entity Recognition | It seeks to locate and classify named entities, such as locations, human names. | token | P | any | WNUT16 NER (Strauss et al., 2016) |
| Part-of-speech Tagging | It is also called grammatical tagging. It identifies words, such as nouns, verbs, adjectives, adverbs, etc. | token | P | any | Ritter11-T-POS (Ritter et al., 2011) |
| Poll Question Generation | It is to generate poll questions for social media posts which will increase user engagement. | context | G | any | dataset collected by (Lu et al., 2021) |
| User Profiling | / | context | G | any | dataset collected by (Liang et al., 2018) |
| Hashtag Generation | / | context | G | any | dataset collected by (Zhang et al., 2021) |
| Post Recommendation | It recommends posts for users according to user engagement history. | context | P | any | dataset collected by (Chen et al., 2012) |
| Rumor Detection | It makes decision that whether a post is rumor or not. | context | P | any | RumourEval (Kochkina et al., 2018) |
| Semantic Analysis | It makes decision that whether a post is positive, negative or neutral. | context | P | any | Semeval2017 Subtask A (Rosenthal et al., 2017) |
| Zero-resource xx (xx is any specific decision task, such as Irony Detection.) | It makes decision on xx without being fine-tuned on task-specific data. | context | P | zero-resource | / |
| Low-resource xx | It makes decision on xx after being fine-tuned on small amount of task-specific data. | context | P | low-resource | / |
| Offensive Language Identification | / | context | P | any | SemEval2019 Task 6 (Zampieri et al., 2019) |
| Irony Detection | / | context | P | any | SemEval2018 Task 3A (Van Hee et al., 2018) |
| Emotion Recognition | / | context | P | any | SemEval-2018 Task 1 (Mohammad et al., 2018) |

the inter-post-comments temperature-scaled cross-entropy loss (PC-TCE). In PC-TCE, post representation $R_P^i$ and corresponding comments set representations $R_C^i$ is a positive sample pair, post representations and comments set representations of the other $N-1$ post-comments pairs are all negative examples where we don't sample negative examples explicitly like (Mikolov et al., 2013; Schroff et al., 2015). We use the cosine distance $sim(u,v) = u^T \cdot v / \|u\| \cdot \|v\|$ to denote the similarity of two vectors, $u$ and $v$, since representations are usually of high dimension. The PC-TCE loss is then defined as:

$$\ell_{PC-TCE} = -\sum_{i=0}^{N-1} log \frac{exp(sim(\tilde{R_P^i}, R_{CI}^i)/\tau)}{\sum_{k=0}^{N-1} \Delta},$$
$$\Delta = \mathbf{1}_{k \neq i} \left[ exp(sim(\tilde{R_P^i}, R_{CI}^k)/\tau) + exp(sim(\tilde{R_P^k}, R_{CI}^i)/\tau) \right], \quad (2)$$

where $\mathbf{1}_{k \neq i} \in \{0,1\}$ is an indicator function evaluating to 1 iff $k \neq i$, $\tau$ is the temperature parameter (Gou et al., 2021), $\tilde{R_P^i}$ is the projection from $R_P^i$ using $f(\cdot)$ since introducing learnable nonlinear transformation improves representation quality in constructive learning (Chen et al., 2020), and $R_{CI}^i$ is the integrated representation of $R_C^i$ obtained by ICCMB.

### 3.4. Joint Learning

The loss function contains three parts corresponding to the three tasks. For MLP, we minimized the cross-entropy loss (CE) between the masked sequence $W^{mask}$ and original sequence $W$,

$$\ell_{MLP} = \sum_{i=0}^{N-1} \sum_{k=0}^{|P^i|} CE(w_k, \hat{w_k}) + \sum_{i=0}^{N-1} \sum_{j=0}^{|C^i|} \sum_{k=0}^{|C_j^i|} CE(w_k, \hat{w_k}), \quad (3)$$

where $w_k$ is the predicted word and $\hat{w_k}$ is the ground truth. For PP, we minimized the CE loss between the predicted predecessor's location of every single comment, $d_k$, and the real value $\hat{d_k}$,

$$\ell_{PP} = \sum_{i=0}^{N-1} \sum_{j=0}^{C^i} \sum_{k=0}^{|C_j^i|} CE(d_k, \hat{d_k}). \quad (4)$$

For PCS-CL, the loss function is $\ell_{PC-TCE}$ which has already been discussed. As a result, the overall loss function for HCCLF is

$$\ell_{HCCLF} = \lambda_1 \cdot \ell_{MLP} + \lambda_2 \cdot \ell_{PP} + \lambda_3 \cdot \ell_{PCS-CL}.$$
$$s.t.\ \lambda_1 + \lambda_2 + \lambda_3 = 1 \quad (5)$$

## 4. Universal Performance Evaluation Benchmark

There are various downstream applications in Social Media Language Understanding. We tend to categorize them from three different perspectives. a. when concentrating on instance granularity, there is the token-level task/context-level task, b. prediction task/generation task, and c. since data-collecting is always human labor demanding, we concentrate on the accessibility of resources, then there is the adequate-resource task/low-resource task/zero-resource task. Besides, it is necessary to know that these taxonomies are not mutually exclusive. For example, a rumor detection task is a context-level prediction task, and, based on real-world scenarios, the corresponding dataset can be anyone of the adequate-resource task/low-resource task/zero-resource task.

To evaluate systematically, we propose to develop a Universal Performance Evaluation Benchmark for social media language, called **Social-UPEB**. It should cover tasks in all the above-mentioned categories. The detail of the proposed evaluation benchmark is illustrated in Table 1.

## 5. Conclusion

In this position paper, we make four contributions. First, we analysis the domain gap between the formal language and the social media language, which hinders present PTMs from being well applied to social media language. Second, we tend to model the social media language with the tree structure which will facilitate in context information modeling and build robust PTMs. Third, based on the above mentioned tree structure, we propose Hierarchical Contextual Constructive Learning Framework (HCCLF) for Social Media Language Understanding. Last, we propose to develop a Universal Performance Evaluation Benchmark for Social Media Language Understanding.

## 6. Bibliographical References


Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637, Long Beach, California, USA. PMLR.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Francesco Barbieri, Luis Espinosa Anke, and José Camacho-Collados. 2022. XLM-T: multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth*

*Language Resources and Evaluation Conference, LREC 2022, 20-25 June 2022*, pages 258–266, Marseille, France. European Language Resources Association.

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1644–1650, Virtual Event. Association for Computational Linguistics.

Farah Benamara, Diana Inkpen, and Maite Taboada. 2018. Introduction to the special issue on language in social media: exploiting discourse and other contextual information. *Computational Linguistics*, 44(4):663–681.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, Virtual Events. Curran Associates, Inc.

Kailong Chen, Tianqi Chen, Guoqing Zheng, Ou Jin, Enpeng Yao, and Yong Yu. 2012. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, pages 1597–1607, Virtual Event. PMLR, PMLR.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Alexandra DeLucia, Shijie Wu, Aaron Mueller, Carlos Aguirre, Philip Resnik, and Mark Dredze. 2022. Bernice: A multilingual pre-trained encoder for twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, December 7-11, 2022*, pages 6191–6205, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Keyang Ding, Jing Li, and Yuji Zhang. 2020. Hashtags, emotions, and comments: a large-scale dataset to understand fine-grained social emotions to online topics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1376–1382, online event. Association for Computational Linguistics.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

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

Maryam Heidari, James H Jones, and Ozlem Uzuner. 2020. Deep contextualized word embedding for text-based online user profiling to detect social bots on twitter. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 480–487, Sorrento, Italy. IEEE.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jing Li, Yan Song, Zhongyu Wei, and Kam-Fai Wong. 2018. A joint model of conversational discourse and latent topics on microblogs. *Computational Linguistics*, 44(4):719–754.

Shangsong Liang, Xiangliang Zhang, Zhaochun Ren, and Evangelos Kanoulas. 2018. Dynamic embeddings for user profiling in twitter. KDD '18, page 1764–1773, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Annie Louis and Shay B Cohen. 2015. Conversation trees: A grammar model for topic structure in forums. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and José Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022 - System Demonstrations, May 22-27, 2022*, pages 251–260, Dublin, Ireland,. Association for Computational Linguistics.

Zexin Lu, Keyang Ding, Yuji Zhang, Jing Li, Baolin Peng, and Lemao Liu. 2021. Engage the public: Poll question generation for social media posts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 29–40, Online. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, November 16-20, 2020*, pages 8968–8979, Virtual Event. Association for Computational Linguistics.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

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

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020a. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14, Virtual Event. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020b. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14, Virtual Event. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 803–818, online event. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534,

Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 815–823, Boston, MA, USA. IEEE.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text, NUT@COLING 2016, Osaka, Japan, December 11, 2016*, pages 138–144. The COLING 2016 Organizing Committee.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017,*, volume 30, pages 5998–6008, Long Beach, CA, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2022. Superb: Speech processing universal performance benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, May 22-27, 2022*, pages 8479–8492, Dublin, Ireland. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Xingshan Zeng, Jing Li, Lu Wang, Zhiming Mao, and Kam-Fai Wong. 2020. Dynamic online conversation recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3331–3341.

Tong Zhang, Yong Liu, Boyang Li, Peixiang Zhong, Chen Zhang, Hao Wang, and Chunyan Miao. 2022a. Toward knowledge-enriched conversational recommendation systems. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 212–217.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2022b. Twhin-bert: A socially-enriched pre-trained language model

for multilingual tweet representations. *CoRR*, abs/2209.07562.

Yuji Zhang, Yubo Zhang, Chunpu Xu, Jing Li, Ziyan Jiang, and Baolin Peng. 2021. #HowYouTagTweets: Learning user hashtagging preferences via personalized topic attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7811–7820, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. *arXiv preprint arXiv:1609.09028*.