Predicting Cross-lingual Trends in Microblogs

Satoshi Akasaki
LY Corporation Research
sakasaki@lycorp.co.jp

Abstract

Trends on microblogs often transcend linguistic boundaries, evolving into global phenomena with significant societal and economic impact. This paper introduces and tackles the novel predictive task of forecasting which microblog trends will cross linguistic boundaries to become popular in other languages, and when. While crucial for proactive global monitoring and marketing, this area has been underexplored. We introduce a methodology to overcome the challenge of cross-lingual trend identification by automatically constructing a dataset using Wikipedia's inter-language links. We then propose a prediction model that leverages a rich feature set, including not only temporal frequency but also microblog content and external knowledge signals from Wikipedia. Our approach significantly outperforms existing trend prediction methods and LLM-based approaches, achieving an improvement of up to 4% in F_1-score, enabling the forecast of cross-lingual trends before they emerge in a new language.

1 Introduction

Microblogs are widely used across the globe to disseminate real-world information rapidly and convey opinions from diverse perspectives. On these platforms, a phenomenon known as a "trend" occurs, where a specific topic gains significant attention over a short period. These trends encompass a wide variety of topics, including local and global news, specific works of art or entertainment, information about individuals or organizations, and memes, influencing both society and individuals.

While trends are typically consumed within their language of origin, some occasionally spread to other languages. For example, the "Ice Bucket Challenge" was a trend that first emerged in English and progressively spread to other languages, ultimately becoming a global social phenomenon. Similarly, public figures such as politicians and artists, products like films and songs, and various events often become trends that transcend linguistic boundaries. Such "cross-lingual trends," which are mentioned not just in one language but across others, are valuable not only for practical applications like enhancing user awareness of global phenomena, but also for sociological studies. Crucially, forecasting such trends before they cross linguistic boundaries enables proactive applications, such as companies launching marketing campaigns in advance or governments implementing relevant policies. While most cross-lingual studies of trends (Juffinger and Lex, 2009; Hale, 2012) focus on retrospectively analyzing trends that have already spread, our work pioneers the forward-looking challenge of prediction. Moreover, existing trend prediction studies have only made forecasts within a single language (Szabo and Huberman, 2010; Asur et al., 2011; Tsur and Rappoport, 2012; Benhardus and Kalita, 2013; Roy et al., 2015; Cheng et al., 2016; Matsuno et al., 2023), failing to address the dynamics of cross-lingual diffusion.

Therefore, we introduce and address the novel task of predicting cross-lingual trends on microblogs. To identify cross-lingual trends, which requires name reconciliation of trend topics across different languages, we propose a method to automatically construct a dataset for this task. Our approach involves treating multilingual Wikipedia entities as trend topics and collecting their microblog posts, considering noise filtering. For robust prediction of cross-lingual trends, we propose using not only the frequency information commonly employed in conventional trend prediction (Shulman et al., 2016) but also a diverse set of features, including the content of microblog posts and external information such as Wikipedia articles. Through experiments using English and Japanese, we demonstrate that our method can predict cross-lingual trends with higher accuracy than existing trend prediction methods and LLM-based approaches.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 530–539

Our main contributions include: (1) formulating the cross-lingual trend prediction task for microblogs; (2) a methodology for automated, filtered dataset construction; and (3) a hybrid feature set integrating temporal, textual, and external signals.

2 Related Work

In the domain of social media, particularly microblogs, numerous attempts have been made to predict trends. However, these studies (Szabo and Huberman, 2010; Asur et al., 2011; Tsur and Rappoport, 2012; Benhardus and Kalita, 2013; Roy et al., 2015; Cheng et al., 2016; Matsuno et al., 2023) have almost exclusively focused on predicting whether a piece of content will become a trend within a single language, thus failing to account for trends that cross linguistic boundaries.

Several studies have addressed various cross (multi)-lingual tasks on social media. To facilitate multilingual content understanding in microblogs, Antypas et al. (2024) and Peng et al. (2022) attempt to transfer an NLP model trained in a source language to a target language. Godavarthy and Fang (2016) focus on retrieving posts in one language that are relevant to posts in another. While these studies operate in a cross-lingual context, their scope is limited to transfer learning or scenarios that assume the pre-existence of relevant posts in the target language. Consequently, their settings are not designed to account for future cross-lingual trends, which is the focus of our research.

Jin et al. (2017) focus on predicting cross-lingual information cascades on microblogs, which tracks the diffusion of individual posts based on user language profiles. In contrast, our work addresses the prediction of macro-level topic trends and leverages external knowledge (Wikipedia) for robust topic reconciliation across languages.

We aim to detect cross-lingual trends at an early stage, before they actually become cross-lingual.

3 Task and Dataset Construction

This section describes the task of predicting cross-lingual trends on microblogs, along with the construction of the dataset. Here, as a first step, we focus on X, a microblog platform with a large user base, and use English and Japanese, the most popular languages in X (Alshaabi et al., 2021).

3.1 Trends and Cross-lingual Trends

On the X platform, topics with high user engagement are designated as "trends." These are determined by X's proprietary algorithm for each language and serve as a de facto indicator of current public interest. Sometimes, a trend in one language can become a trend in another. We refer to such trends as "cross-lingual trends," indicating that they have gained global interest across linguistic barriers. This study aims to detect these trends at an early stage before they become cross-linguistic.

Here, since X's official trending algorithm[1] is proprietary and its data is not publicly available, we must first define our own concept of a "trend." Typically, a trend is defined not merely by high frequency of mentions or localized buzz, but as a topic that garners a significant volume of posts within a concentrated period. Drawing on existing studies (Vlachos et al., 2004; Graus et al., 2018), we define a trend based on frequency information from archival X data as follows:

Trends.

$$= \{w_t | MA_w(t) > \gamma, Count(w_t) >= \theta\} \quad (1)$$

$$\gamma = mean(MA_w) + 1.5 * std(MA_w) \quad (2)$$

Here, w represents a candidate trend phrase, t is a timestamp. Let $Count(w_t)$ be the mention frequency on t. MA_w is the 7-day moving average of $Count(w_t)$. γ is a threshold used to detect a "burst," which is a sudden increase in mentions for w. According to this definition, w is considered a trend if it experiences a burst and its frequency exceeds a threshold θ. θ, designed to prune minor trend candidates, is set based on the top 20% of frequencies observed for historical trends in the target language. The remaining parameters are adopted from Vlachos et al. (2004).

A cross-lingual trend is then defined as follows:

Cross-lingual trends. *A trend that originates in one language and subsequently becomes a trend in other languages.*

These cross-lingual trends often involve deeper dynamics, including cultural translatability, global relevance, media coverage, and platform-specific propagation patterns. This requires a different approach than traditional trend forecasting.

[1] https://help.x.com/en/using-x/x-trending-faqs

3.2 Task Definition

The objective of this research is to enable proactive responses to cross-lingual trends. This requires (1) identifying potential cross-lingual trends in advance to facilitate timely actions such as public opinion monitoring and marketing preparation, and (2) estimating the approximate timing of the trend becoming cross-lingual to enable strategic planning. Therefore, we formally define our tasks as follows: (a) predict whether a trend that has emerged in language A will also become a trend in language B, and (b) if so, estimate its approximate timeframe.

A key challenge is predicting a trend's future emergence in another language, which distinguishes this task from conventional trend prediction within a monolingual language. Consequently, relying solely on features like time-series frequency, as recommended in existing studies (Shulman et al., 2016), is insufficient. To balance the need for early prediction with the requirement of sufficient data for robust modeling, we adopt the following practical setting: *predictions are made using both language data available within one month of the trend's initial emergence.* This setting is designed to capture a broad range of early signals from both languages without unduly delaying the prediction.

3.3 Dataset Construction

Constructing a dataset for predicting cross-lingual trends presents several challenges. First, to collect cross-lingual trends, name reconciliation of identical trends across languages is necessary. However, trend data with such reconciliation does not exist, and collecting it from scratch is difficult. Additionally, microblogs are noisy; trends of X often include uninformative items like greetings or multiple variations of the same topic. The microblog posts themselves are also inherently noisy.

To address these issues, we propose treating entities from Wikipedia as trends. This approach was adopted as a practical solution to the formidable challenge of cross-lingual name reconciliation, given the lack of publicly available, officially labeled trend data. Wikipedia entities have inter-language links, which can be followed to obtain the entity's name in each language, thereby systematically resolving the name reconciliation problem for any language pair. Moreover, entities in Wikipedia cover a wide range of topics, from creative works to memes, [2] are registered with their official names, and are required to meet a standard of notability. This approach allows us to build a "clean and useful" benchmark dataset, which is essential for pioneering research on this new problem. We construct the dataset by treating these Wikipedia entities as trends and collecting their corresponding microblog posts. Here, to fully leverage the multilingual environment, we collect posts in both languages for each target trend. The specific steps of this procedure are detailed below.

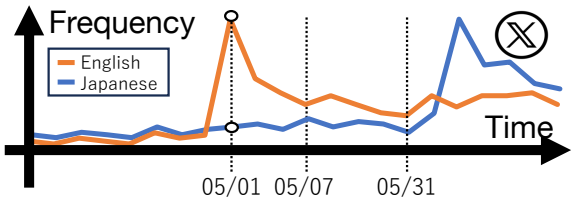

Figure 1: Frequency plot of Trend A on X. For a trend A that first emerges in English, we extract posts from both languages within the three time periods separated by the dotted lines, including its point of emergence (white dot). This data is then used to predict whether A will subsequently become a trend in Japanese.

First, using the Wikipedia API,[3] we extract entities newly registered between 2012 and 2023 that have inter-language links between Japanese and English. During this process, homonymous entities can be included, as posts about different topics that share the same entity name may be collected. To mitigate this, for each language, we use X posts from before 2011 to identify and remove entities that had already appeared, treating them as pre-existing homonyms. Next, for each entity, we count the time-series frequency of the entity's mentions on X posts for both languages, and we determine if a trend has occurred using the definition (§ 3.1). For the entity identified as a trend in at least one language, we extract 100 random posts from each language over three distinct periods, which are defined relative to the trend's point of emergence in its initial language: the first day of the trend, the first week, and the first month (Figure 1). At this time, since microblog trends are susceptible to contamination from spam or keyword stuffing, direct collection can introduce significant noise. As a simple countermeasure, we prioritize collecting posts with a high number of retweets, leveraging the assumption that highly retweeted

[2] https://en.wikipedia.org/wiki/Category:Popular_culture_language

[3] https://www.mediawiki.org/wiki/API:Main_page

	Lang.	#Trends	#Posts
Cross-lingual	En→Ja	1,210	110.4k
	Ja→En	1,114	103.8k
Mono-lingual	En	4,927	479.4k
	Ja	4,582	438.2k

Type	#Trends
PRODUCT	1,116
PERSON	651
GROUP	254
EVENT	189
LOCATION	114

Table 1: Collected trends: The left shows both cross-lingual and mono-lingual trends. The right shows the results of classifying cross-lingual trends into 5 types. 'En→Ja' denotes trends that first appeared in English and later in Japanese, and vice versa.

posts are more likely to be informative and reflect genuine user interest. Finally, instances identified as trends in both languages are labeled as positive examples (Cross-lingual), while the rest are labeled as negative examples (Mono-lingual).

The results of data collection are summarized in Table 1 (left). Here, we partition the data based on the language in which the trend first emerged. As expected, the proportion of cross-lingual trends is small compared to mono-lingual ones, and a majority of cross-lingual trends originate in English. To further investigate the nature of these trends, we classified each trend into one of the 5 named entity types defined by Sekine et al. (2002), by feeding its Wikipedia summary into GPT-4o.[4] The results are shown in Table 1 (right). Notably, PRODUCT-type trends (*e.g.,* creative works, products) account for the majority. The next most frequent are PERSON-type trends, indicating that these two types are more likely to become cross-lingual. Table 2 classifies cross-lingual trends by the time it took for them to become cross-lingual, showing a broad period from a few days to several years. Recognizing that these durations often align with meaningful intervals (*e.g.,* weeks, months, and years), we analyze the trends across four distinct periods:

$(0 \leq day \leq 7)$**:** This is the largest category, featuring highly anticipated works like "Avengers: Endgame" and breaking news topics such as the "Mt. Gox". Such high-profile trends tend to propagate rapidly across languages.

$(7 < day \leq 31)$**:** This category includes minor but globally significant topics like the "Panama Papers" and "Jeffrey Epstein," which attract public attention gradually rather than abruptly.

$(31 < day \leq 365)$**:** We see examples like the "New Nintendo 3DS," where products or media intended for a multinational audience are announced in one country and then gain presence in others over an extended period.

$(365 < day)$**:** This group contains instances like "Lyft" or the "Temple of Bel," which were initially confined to one country but later gained cross-lingual presence due to sudden news events or their introduction into other markets.

This analysis reveals the diverse nature and timing patterns of cross-lingual trends, highlighting the need to leverage not only temporal patterns but also the contextual information embedded in trend-related content for effective prediction. We frame the estimation of the cross-lingual trend duration as a classification problem over these four categories.

4 Experiments

We try to predict cross-lingual trends by developing classifiers using the constructed dataset.

4.1 Features

To predict cross-lingual trends, we develop a comprehensive feature set that extends beyond conventional temporal metrics. While temporal features are crucial for general trend prediction (Shulman et al. (2016)), predicting long-term cross-linguistic solely based on initial temporal patterns is challenging. We hypothesize that textual content from microblog posts and user profiles, along with external signals from Wikipedia, can provide vital clues for cross-lingual trends. Textual features may capture cross-lingual clues, while Wikipedia data can provide more detailed knowledge about the trend and its demand beyond the microblog sphere. Here, to effectively handle text from multiple languages, we adopt multi-lingual sentence embeddings (Artetxe and Schwenk, 2019). This allows for the sharing of features across different language pairs by converting the text into a universal vector representation.

Our feature set is extracted from three distinct periods defined in § 3.3 for both languages. This multi-period approach captures both immediate reactions and evolving signals. Here, to handle multilingual text, we adopt multilingual sentence embeddings. The features are categorized as follows:

Temporal Frequency: The temporal frequency of posts associated with the trend. For each trend, we calculate and use the maximum gradient (*i.e.,* the steepest rate of change in post frequency over time), average gradient, maximum frequency, and average

[4] https://chatgpt.com/

Days to cross-lingual	#Trends	Examples of cross-lingual trends: trend [TYPE]
$0 \leq day \leq 7$	743	Cameron Boyce [PERSON], Mt. Gox [GROUP], Avengers: Endgame [PROD.]
$7 < day \leq 31$	474	Panama Papers [OTHER], Istanbul Airport [LOC.], Jeffrey Epstein [PERSON]
$31 < day \leq 365$	570	New Nintendo 3DS [PROD.], Luka Modrić [PERSON], Blade Runner [PROD.]
$365 < day$	537	Lyft [GROUP], World Emoji Day [EVENT], Temple of Bel [LOC.]

Table 2: Trends categorized by the number of days until becoming cross-lingual.

gradient before and after the time of maximum frequency for the target period.

Posts: Textual content of posts in the trend. To aggregate posts, for each trend, we obtain sentence embeddings for each post and use the average of these embeddings in each period.

User Profiles: User profiles of the trend. For each trend, we obtain sentence embeddings of the profiles of the users, who posted about the trend, and use the average of these embeddings in each period.

Wikipedia Lead Sentences: Introductory section of the Wikipedia article. For each trend, we acquire the lead sentences of the corresponding article and then compute their sentence embeddings. We average them to create the feature vector.

Wikipedia Pageviews: The pageview frequency of the Wikipedia article.[5] For each trend, we extract the pageviews of the corresponding article and use the maximum gradient, average gradient, maximum frequency, and average gradient before and after the time of maximum frequency for the target period.

4.2 Comparison Methods

To evaluate the efficacy of our proposed approach, we compare its performance against several baseline methods, each representing a distinct strategy:

Temporal: This baseline utilizes only the frequency of posts, a common approach in trend prediction. Specifically, daily post frequencies from one month prior up to the prediction date are fed into a Transformer model (Vaswani, 2017) to predict the frequency on the potential trend date. A trend is classified as cross-lingual if its predicted frequency surpasses a predefined threshold.

Cheng: The method by Cheng et al. (2016), originally designed for predicting recurring cascades of (often non-textual) content, employs features such as temporal patterns, user demographics, and network structure (details in Appendix A). Crucially, it does not incorporate textual content from posts or external data like Wikipedia.

GPT-4o: We employ GPT-4o, a state-of-the-art large language model, in a few-shot classification setting. The model is provided with prompts describing the task, the lead sentence of the trend in Wikipedia, and a small number of training examples (see Appendix B) to make a prediction.

Proposed: Our full proposed method, which trains a classification model utilizing the complete hybrid feature set described in § 4.1. Additionally, since this relies solely on frequency information and sentence embeddings as features, data from both languages can be used directly for training. We thus refer to the model trained on monolingual data as **Proposed (mono.)** and the one trained on data from both languages as **Proposed (multi.)**.

4.3 Settings

For sentence embeddings, we use LaBSE (Feng et al., 2022), a large-scale, pre-trained model capable of handling over 100 languages.[6]

For the classification models, we used gradient boosting trees, which combine decision trees and ensemble learning (Chen and Guestrin, 2016). We implemented the model using Python and LightGBM.[7] We optimized hyperparameters with Optuna's[8] LightGBM Tuner, which utilizes Bayesian optimization to determine the parameters automatically. We show hyperparameters in Appendix A.

We treat the task of predicting whether a trend will become cross-lingual as a binary classification, and the task of estimating the period until becoming cross-lingual as a 4-class classification, predicting one of the four periods defined in § 3.3.[9] We perform the task in both directions: predicting whether trends that first emerge in English will also become trends in Japanese, and vice versa. We used the constructed data up to 2021 as the training data and data from 2022 as the evaluation data (see Table 4 for label distribution). Since the number of negative examples was higher than positive examples

[5] https://pageviews.wmcloud.org/

[6] https://tfhub.dev/google/LaBSE/

[7] https://lightgbm.readthedocs.io/en/stable/

[8] https://optuna.org/

[9] For period prediction, we train the model using only the positive examples within the dataset.

	$\mathbf{F}_1$ (En→Ja)	$\mathbf{F}_1$ (Ja→En)
Temporal	55.83	51.92
Cheng (d)	63.76	63.08
GPT-4o (d)	66.42	65.82
Proposed (mono.) (d)	67.25	67.61
Proposed (multi.) (d)	**70.43**	**68.72**
Cheng (d, w)	68.14	68.82
GPT-4o (d, w)	69.17	70.45
Proposed (mono.) (d, w)	70.19	72.31
Proposed (multi.) (d, w)	**72.87**	**73.15**
Cheng (d, w, m)	72.11	72.31
GPT-4o (d, w, m)	71.74	72.02
Proposed (mono.) (d, w, m)	74.22	73.62
Proposed (multi.) (d, w, m)	**76.12**	**75.71**

Table 3: Results of cross-lingual trend prediction. All the scores are macro-averaged to account for class imbalance. d, w, and m represent features extracted on the day of emergence, a week later, and a month later, respectively. **Proposed (multi.)** outperforms all comparisons significantly (measured by Wilcoxon rank-sum test with p-value < 0.05).

En→Ja	**P**	**R**	$\mathbf{F}_1$	**#Trends**
Mono-lingual	89.74	90.56	90.15	975
Cross-lingual	63.20	61.00	62.08	259

Ja→En	**P**	**R**	$\mathbf{F}_1$	**#Trends**
Mono-lingual	90.11	91.01	90.56	901
Cross-lingual	62.15	59.64	60.87	223

Table 4: Detailed results of cross-lingual trend prediction for **Proposed (multi.)** (d, w, m)

	$\mathbf{F}_1$ (En→Ja)	$\mathbf{F}_1$ (Ja→En)
All	**76.12**	**75.71**
– **Posts**	70.92	70.34
– **Temporal Frequency**	70.01	69.41
– **User Profiles**	74.48	74.56
– **Wikipedia Lead Sentences**	72.29	73.01
– **Wikipedia Pageviews**	72.97	71.91

Table 5: Ablation test of **Proposed (multi.)** (d, w, m). "–" indicates the removal of only that feature.

in the training data, we used the 'is_unbalance' option in LightGBM to give more weight to positive examples during training. We used 10% of the training data as development data and applied the model with the lowest loss on the development data to the evaluation data. The scores were obtained by averaging the results over 10 runs. While we utilize information from the month following the date the target trend emerged, there are instances in the test data where the trend becomes cross-lingual within that month. For such cases, predictions are made without using the week or month's information to prevent data leakage.

4.4 Results

Table 3 presents the results for cross-lingual trend prediction on the evaluation data. As the period of features is extended, the F_1-score improves across all methods. This suggests that leveraging longer-term data allows for a more accurate capture of a trend's persistence and growing interest. **Proposed** methods achieve significantly higher performance than the baselines in both prediction directions. This demonstrates that our hybrid feature set, which integrates traditional temporal information with textual content from microblogs and external knowledge from Wikipedia, is highly effective for predicting cross-lingual trends. Moreover, the superior performance of **Proposed (multi.)** over **Proposed (mono.)** underscores the significant benefit of our feature design. This improvement is primarily driven by the use of multi-lingual sentence embeddings, which map textual features from both English and Japanese into a shared vector space. This allows the **Proposed (multi.)** model to be trained on a combined dataset of trends originating from both languages, effectively doubling the training data available. By learning from a larger and more diverse set of examples, the model can capture more robust and generalizable patterns of how trends propagate across languages, leading to its enhanced predictive accuracy. The low performance of **Temporal** and **Cheng** indicates that commonly used features alone, such as temporal information, are insufficient to capture cross-lingual trends. Despite being provided with rich text information, **GPT-4o** is outperformed by the proposed method. This highlights the importance of incorporating diverse signals, such as post momentum and external interest, which a standalone LLM may not capture (Xu et al., 2024).

Table 4 provides a breakdown of the prediction results for **Proposed (multi.)** (d, w, m) by label. While the performance for mono-lingual trends is high, the precision for cross-lingual trends is relatively low. This shows that the method tends to over-predict positive instances, highlighting the inherent difficulty of separating cross-lingual and mono-lingual trends. The imbalanced nature of the evaluation data, where cross-lingual trends are a minority, also contributes to this challenge.

Table 5 shows the ablation test to evaluate the impact of each feature. **Posts** and **Temporal Frequency** contribute the most significantly, indicating that a trend's linguistic content and its momentum are central to predicting its cross-lingual poten-

	$\mathbf{F}_1$ (En→Ja)	$\mathbf{F}_1$ (Ja→En)
$0 \leq day \leq 7$	69.32	65.26
$7 < day \leq 31$	44.83	51.38
$31 < day \leq 365$	58.01	51.22
$365 < day$	50.59	61.79

Table 6: Results of estimating period until the trend becomes cross-lingual by **Proposed (multi.)** (d, w, m)

tial. Features derived from **Wikipedia** (especially pageviews) also make a notable contribution, underscoring that capturing public interest outside the microblog sphere is crucial for improving performance. In contrast, the contribution of **User Profiles** is limited, which may suggest that the current method of aggregating profile embeddings lacks sufficient discriminative power, or that user interest is already captured by the content of their posts.

Table 6 shows the results of classifying the period until a trend becomes cross-lingual. The overall modest F_1-scores indicate that predicting the exact timing is a highly challenging task. The $0 \leq day \leq 7$ category achieves the highest performance, likely because trends in this group, such as globally released films ("Avengers: Endgame"), possess strong and clear signals for rapid propagation. Conversely, trends that take months or years to become cross-lingual are often influenced by subsequent events that are difficult to predict from initial information, resulting in lower accuracy. Addressing these hard-to-predict, long-tail trends remains a key challenge for future work.

We show a brief qualitative analysis. Our model predicted not only easily predictable cross-lingual trends like PRODUCT-type but also news events like the "Wagner Group." These topics drew large attention on Wikipedia in parallel with microblogs, indicating that our features captured broader signals of public interest. On the other hand, there were some false positives, like "Takopi's Original Sin," which became a social phenomenon in Japanese but didn't gain much traction in English. Such examples highlight the need for a better grasp of the target language's culture.

5 Conclusion

We introduced the novel task of predicting cross-lingual trends on microblogs. To enable this, we proposed a methodology to automatically construct a dataset using Wikipedia's inter-language links. Our prediction model leverages a hybrid feature set that integrates temporal dynamics, textual content, and external knowledge signals from Wikipedia. Experiments demonstrated that our approach outperforms existing methods, highlighting that combining diverse signals is key to capturing the complex dynamics of cross-lingual trends. From an industrial perspective, this task holds promise for advancing global marketing and content strategies.

Since this was an initial attempt, our current study has only been validated using Japanese and English. We plan to apply our method to other language pairs to investigate the prediction accuracy and the influence of linguistic differences. We also plan to apply the method to our on-site microblog search service,[10] and release the dataset to facilitate further research.

6 Limitations

While we present a novel approach to predicting cross-lingual trends, we acknowledge several limitations that offer avenues for future research.

Limited Language Scope: Our experiments are confined to the English-Japanese language pair. We acknowledge that the dynamics of trend propagation can be heavily influenced by the cultural, economic, and political proximity between language communities. Consequently, our findings may not be directly generalizable to other language pairs with different characteristics. However, our choice to begin with English and Japanese was a deliberate starting point for this novel task. As noted, these two languages represent two of the largest and most active language communities on the X platform, providing a robust and challenging testbed for our initial investigation. The cultural and economic distance between them also makes for a non-trivial case study. Crucially, our proposed methodology, which leverages a language-agnostic sentence encoder (LaBSE), is designed to be extensible. This architecture was chosen specifically to facilitate future expansion. Future work should expand the dataset to include a more diverse range of languages (*e.g.,* language pairs with different cultural proximities like English-Spanish or Chinese-Japanese) to investigate both universal and language-pair-specific factors that govern cross-lingual trend propagation.

Dependence on Wikipedia for Trend Identification: Our methodology relies on Wikipedia entities to identify and reconcile trends across languages. This approach, while effective for system-

[10] https://search.yahoo.co.jp/realtime

atically collecting clean and notable topics, has an inherent limitation: it cannot capture ephemeral or informal trends, such as emergent memes, slang, or grassroots events that have not yet been documented on Wikipedia. We adopted this method as a practical solution to the formidable challenge of cross-lingual name reconciliation, given the lack of publicly available, officially labeled trend data. To address this, future research could explore methods for trend identification that are independent of Wikipedia, such as cross-lingual topic modeling or clustering techniques, which could capture a broader spectrum of trends.

Difficulty in Predicting Long-Term Trends: As shown in our results, the model's performance decreases significantly when predicting "long-tail" trends that take months or even years to cross linguistic boundaries. This is an inherent challenge of the task, as the eventual spread of such trends is often triggered by unforeseen external events (*e.g.,* a sudden news report about a company, a product's delayed international release) that cannot be inferred from the initial data. While our model demonstrates strong performance for short-term trends, which holds significant practical value for applications like real-time marketing and public opinion monitoring, improving long-term prediction remains a key challenge. Future models could incorporate dynamic updates, continuously integrating new information as it becomes available to revise predictions over time.

Interpretability of the Prediction Model: Our proposed model, which uses a gradient boosting framework, achieves high predictive accuracy but offers limited interpretability. While our ablation study (Table 5) reveals the relative importance of different feature categories, it does not explain the specific reasons behind a prediction for an individual trend. Understanding why a trend is predicted to become cross-lingual is crucial for gaining deeper sociolinguistic insights into the mechanisms of information diffusion. Future work could employ model-agnostic explanation techniques, such as SHAP or LIME, to analyze individual predictions and uncover the complex interplay of features that drive a trend's cross-lingual potential.

7 Ethics Statement

We collected the dataset using X's official API and in compliance with X's terms of use. We plan to distribute sentence embeddings or IDs of the posts used in the experiments, and we ensure that their redistribution complies with X's developer policy. Researchers cannot collect deleted posts or posts of private users, thus protecting user privacy.

References

Thayer Alshaabi, David Rushing Dewhurst, Joshua R. Minot, Michael V. Arnold, Jane L. Adams, Christopher M. Danforth, and Peter Sheridan Dodds. 2021. The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *EPJ Data Science*, 10(1).

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, and Jose Camacho-Collados. 2024. Multilingual topic classification in X: Dataset and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Sitaram Asur, Bernardo A Huberman, Gabor Szabo, and Chunyan Wang. 2011. Trends in social media: Persistence and decay. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 5, pages 434–437.

James Benhardus and Jugal Kalita. 2013. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (KDD)*, pages 785–794.

Justin Cheng, Lada A. Adamic, Jon M. Kleinberg, and Jure Leskovec. 2016. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, page 671–681.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891.

Archana Godavarthy and Yi Fang. 2016. Cross-language microblog retrieval using latent semantic modeling. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 303–306.

David Graus, Daan Odijk, and Maarten de Rijke. 2018. The birth of collective memories: Analyzing emerging entities in text streams. *Journal of the Association for Information Science and Technology*, 69(6):773–786.

Scott A Hale. 2012. Net increase? cross-lingual linking in the blogosphere. *Journal of Computer-Mediated Communication*, 17(2):135–151.

Hongshan Jin, Masashi Toyoda, and Naoki Yoshinaga. 2017. Can cross-lingual information cascades be predicted on twitter? In *International conference on social informatics (SocInfo)*, pages 457–472. Springer.

Andreas Juffinger and Elisabeth Lex. 2009. Cross-language blog mining and trend visualisation. In *Proceedings of the 18th international conference on World wide web (WWW)*, pages 1149–1150.

Shogo Matsuno, Sakae Mizuki, and Takeshi Sakaki. 2023. Construction of evaluation datasets for trend forecasting studies. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, pages 1041–1051.

Hao Peng, Ruitong Zhang, Shaoning Li, Yuwei Cao, Shirui Pan, and Philip S Yu. 2022. Reinforced, incremental and cross-lingual event detection from social messages. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):980–998.

Suman Deb Roy, Gilad Lotan, and Wenjun Zeng. 2015. The attention automaton: Sensing collective user interests in social network communities. *IEEE transactions on network science and engineering*, 2(1):40–52.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)*.

Benjamin Shulman, Amit Sharma, and Dan Cosley. 2016. Predictability of popularity: Gaps between prediction and understanding. In *Proceedings of the international AAAI conference on web and social media (ICWSM)*, pages 348–357.

Gabor Szabo and Bernardo A Huberman. 2010. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88.

Oren Tsur and Ari Rappoport. 2012. What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM)*, pages 643–652.

A Vaswani. 2017. Attention is all you need. *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*.

Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 131–142.

Hanzi Xu, Renze Lou, Jiangshu Du, Vahid Mahzoon, Elmira Talebianaraki, Zhuoan Zhou, Elizabeth Garrison, Slobodan Vucetic, and Wenpeng Yin. 2024. Llms' classification performance is overclaimed. *Preprint*, arXiv:2406.16203.

A Hyperparameters

For **Temporal**, we use Transformer (Vaswani, 2017) as a prediction model. We set the head size to 256, the number of heads to 4, the size of the hidden layer in the feed-forward network to 32, the number of transformer blocks to 3, the number of epochs to 200, the learning rate to 0.001, and the batch size to 64. Additionally, we configured a dropout layer with a probability of 0.2 before the final layer.

For **Cheng** (Cheng et al., 2016), note that they target Facebook, which includes features that may not be applicable to microblogs. In our work, we used **Temporal Frequency** as temporal features and the distribution of age, gender, and country of users as demographic features. For network features, we used the number of users, the number of follows per user, and the number of followers per user, and we used the number of reposts and the distribution of reposts as multiple-copy features.

B Settings of GPT-4o

Here, we describe the settings of GPT-4o used in the paper. We use GPT-4o on 1 June 2024; the temperature is set to 0. For each trend, we give a maximum of 100 posts in the dataset.

The following prompt is used for cross-lingual trend prediction (§ 4). For a few-shot classification, we give 10 examples of training data.

```
*We provide a trend of X and
its corresponding introduction of
the Wikipedia article, posts, and
we would like you to assign a
specific label to them.

*For each trend, the introduction
is given, followed by the group
of posts separated by commas.

*For each trend, read the input
and determine whether the trend
will become a trend in another
language.

*Output a label of 1 if it will
become a cross-lingual trend, and
0 if it will not.
```

*Trend: [TREND]

*Posts: [INTRO AND POSTS]

*Answer: [LABEL]