Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Construction of Evaluation Datasets for Trend Forecasting Studies

Shogo Matsuno*[2], Sakae Mizuki*[1], Takeshi Sakaki [1]

[1] Hottto Link, inc.
[2] Gunma University
s.matsuno@gunma-u.ac.jp, s.mizuki@hottolink.co.jp, t.sakaki@hottolink.co.jp

Abstract

In this study, we discuss issues in the traditional evaluation norms of trend forecasts, outline a suitable evaluation method, propose an evaluation dataset construction procedure, and publish Trend Dataset: the dataset we have created. As trend predictions often yield economic benefits, trend forecasting studies have been widely conducted. However, a consistent and systematic evaluation protocol has yet to be adopted. We consider that the desired evaluation method would address the performance of predicting which entity will trend, when a trend occurs, and how much it will trend based on a reliable indicator of the general public's recognition as a gold standard. Accordingly, we propose a dataset construction method that includes annotations for trending status (trending or non-trending), degree of trending (how well it is recognized), and the trend period corresponding to a surge in recognition rate. The proposed method uses questionnaire-based recognition rates interpolated using Internet search volume, enabling trend period annotation on a weekly timescale. The main novelty is that we survey when the respondents recognize the entities that are highly likely to have trended and those that haven't. This procedure enables a balanced collection of both trending and non-trending entities. We constructed the dataset and verified its quality. We confirmed that the interests of entities estimated using Wikipedia information enables the efficient collection of trending entities a priori. We also confirmed that the Internet search volume agrees with public recognition rate among trending entities.

1 Introduction

The prediction of trends is a topic of interest across a wide range of industries, as it often yields economic benefits. In product development, for example, a company can get ahead of its competition by developing new services using information obtained from trend forecasts, resulting in higher sales and consumer satisfaction (Forslund and Jonsson 2007; Yu and Kak 2012). In production and inventory management, efficient production, manufacturing, and inventory planning based on sales forecasts that utilize predictions of trends can reduce business risks and costs, and contribute to the realization of a sustainable society through waste reduction (Dogan and Birant 2021; Pournader et al. 2021; Murray, Agard, and Barajas 2018).

Although a lot of studies on trend forecasting have been conducted, the definition, gold standards, and quantification methodologies for social trends are diverse and depend on their target domains[1] (Chambers, Mullick, and Smith 1971). Given these considerations, we define a trending phenomenon as one in which the public recognition rate of a specific entity[2] increases rapidly in a short period. The recognition rate is defined as the percentage of consumers in society as a whole or within a sample survey who recognize a particular entity. Based on this definition, trend forecasting can be defined as the task of predicting a rapid increase in the recognition rate over time within a population. With a constant supply of new entities, the extent to which they attain popularity within society can vary greatly. Consequently, a protocol for evaluating a trend forecasting method should encompass an assessment of its ability to accurately predict the trending status, degree of trending, and trend duration.

However, to the best of our knowledge, no consistent and systematic evaluation protocol for trend forecasting methods has been developed. For example, some studies evaluated using trended entities alone without considering the ability to distinguish between trending and non-trending entities (Li et al. 2017). Other studies focused on evaluating the classification performance of trending status while ignoring the prediction of periods corresponding to those trends (Bandari, Asur, and Huberman 2012; Yu, Asur, and Huberman 2015). Furthermore, the domains and entities subject to evaluation are not standardized, making it difficult to compare the generality and performance of the proposed methods. Moreover, some studies employed the volume of Internet activity as a proxy indicator of the public recognition rate, which may not always be a valid assumption (Rousidis, Koukaras, and Tjortjis 2020). Consider the case of using the Internet search volume of the word, or the word frequencies on social media posts, as proxy indicators. Although these measures represent public interests, they do not necessarily correlate with public recognition rates, as they may be affected by seasonal

*These authors contributed equally to this work

[1] The term "domain" refers to a specific area that is used to categorize the types of entities.

[2] The term "entity" refers to a distinct object or a named entity that is uniquely recognized and identified, such as a person, a product, or a geographic location.

Entity name	Alina Zagitova	Düsseldorf Airport Terminal station	Back to the Future
Entity name (Ja)	アリーナ・ザギトワ	デュッセルドルフ空港ターミナル駅	バック・トゥ・ザ・フューチャー
Domain	Person/Group-Athlete	Geography–City/Region/Landmark	Art/Content–Movie
Domain (Ja)	人物・グループ-スポーツ選手	地理-都市・地域・ランドマーク	創作物-映画
Interest pattern	`positive`	`negative`	`negative-popular`
Recognition rate: survey start time	0.047	0.044	0.896
Recognition rate: survey end time	0.837	0.097	0.945
Trending status	Trending	Non-trending	Non-trending
Degree of trending	0.790	0.053	0.049
Trend period: start date	October 29, 2017	—	—
Trend period: end date	April 29, 2018	—	—

Table 1: Examples of entities present in the Trend Dataset, which is created through the proposed method. Trending status and degree of trending are annotated based on a sample survey conducted in Japan. The trend period is annotated based on integrating sample survey and Internet search volume. Wikipedia article metadata is used to assign domain and interest pattern, which are then employed to collect trending and non-trending entities from diverse domains in a balanced manner.

or news-driven changes in public interests.

Ultimately, the following four issues can be identified while scrutinizing existing trend forecasting studies: usage of both trending and non-trending entities, trend period forecasting, standardization of domains and entities subject to evaluation, and adoption of valid public recognition rate indicators. Accordingly, evaluation datasets must be annotated[3] based on the time series of public recognition rates collected for both trending and non-trending entities, with a maximal temporal resolution and range of domains. However, according to our survey, no existing datasets conform to these specifications, as constructing such a dataset is laborious and costly. Considering that trending entities cannot be known in advance and that trending entities represent a small fraction of the total, a sufficiently large set of entities must be surveyed. Furthermore, frequent sampling is necessary to achieve high temporal resolution. These facts suggest the need to use different methods than conventional social surveys.

Therefore, we propose an efficient procedure that collects both trending and non-trending entities and measures the weekly recognition rate time series. There are two key ideas. Firstly, a sample survey of the public recognition rate is conducted similarly to a retrospective cohort study. Specifically, we collect the entities to be surveyed based on the interests of entities estimated using Wikipedia information, and then we survey the year when the general public recognizes entities. Secondly, the Internet search volume is used to interpolate the survey-based yearly recognition rate time series on a weekly timescale.

The key ideas mentioned above solve two problems associated with the construction of trend forecast evaluation datasets. The first idea ensures a balanced collection of entities without requiring the survey of many entities. The second idea enables a weekly timescale for trend periods without doing frequent surveys. The contributions of the methodology and the dataset constructed in this study to trend forecasting research are as follows:

- Trending status is determined using the social survey-based recognition rate, which yields more faithful gold standard than proxy indicators such as Internet search volume.
- Both trending and non-trending entities are collected, enabling the evaluation of the classification accuracy of the trending status.
- The weekly recognition rate time series is measured, allowing for an evaluation of the trend period's prediction accuracy.
- Entities from a wide range of domains are collected, suitable for a standardized evaluation of the trend forecasting methods.

We employed the proposed method to collect entities, measure recognition rates, and annotate trend attributes. We make the compiled dataset, Trend Dataset, publicly available. Table 1 shows examples of entities in the dataset. We also analyzed the relationship between interests estimated using Wikipedia information, crowdsourcing-based public recognition rates, and Internet search volumes to confirm the validity of the key ideas. We confirmed that, overall, analysis results are consistent with these ideas.

2 Related Works

One of the widely accepted methods for trend forecasting is the utilization of the volume of internet activity as an explanatory variable. For example, Choi et al. (2012) proposed an algorithm to predict time-series data of various economic indicators for the sales industry using the internet search volume of some keywords. Skenderi et al. (2021) conducted sales trend prediction in the fashion industry. Sayyad et al. (2009) predicted trends in events using the frequency of occurrence for a set of keywords that refer to an event and story as explanatory variables. Jang et al. (2021) predicted real-world outbreaks of illness from the occurrence of the word "cold" among online sources.

However, to the best of our knowledge, there exist no studies that have systematically defined a consistent evaluation methodology across various domains by examining the

[3]The term "annotation" refers to the process of assigning labels or descriptive information to raw data.

Dataset	Trending and non-trending	Target domains	Trending indicators	Trending status	Degree of trending	Trend period
Trend Dataset (Ours)	**both**	**21 (Table 3)**	**social survey**	**yes**	**yes**	**yes**
Bandari+ (2012)	**both**	**31 (news categories)**	tweet counts	**yes**	**yes**	no
Choi & Varian (2012)	only trending	automotive, home, travel	product sales, visitor volume	no	**yes**	no
Skenderi+ (2021)	**both**	fashion	product sales	no	**yes**	no
Jang+ (2021)	only trending	infection	**epidemiological survey**	no	**yes**	no

Table 2: A comparative analysis between the existing datasets and the Trend Dataset from six perspectives: coverage of both trending and non-trending entities, target domains, indicators used to quantify public recognition, and the availability of trending status, degree of trending, and trend period.

evaluation metrics and targeted domains (such as product categories and industries) employed in previous research. Many studies of evaluated trend forecasting methods based on weighing the differences between predicted and actual values. For example, the aforementioned study conducted by Jang et al. (2021) predicted the weekly number of reported cases and evaluated its performance using the root-mean-squared deviation (RMSE) between the predicted and actual results. Tsur et al. (2012) predicted the proliferation of memes on social media, using the number of tweets containing a specific hashtag as a prediction target to evaluate the mean squared error (MSE) and correlation between predictions and actual results. Furthermore, a study by Murray et al. (2018) on supply chain demand forecasting used the mean absolute percent error (MAPE) and RMSE as quantitative measures between forecasts. However, indices such as RMSE and correlation are measures for evaluating time-series forecasting models and are not necessarily useful for evaluating trend phenomena. Rather than predicting indicators based on public recognition rates, some evaluation methods predict proxy indicators based on interest, such as Internet search volume. A study by Suman et al. (2015) on Twitter trend prediction used community volatility (frequency of change in trending topics) and affinity as proxy measures. Another study by Cheng et al. (2016) on the diffusion phenomenon of Facebook posts considered the recurrence of rapid spreading and their extent as proxy indicators. The variable nature of proxy indicators makes cross-study comparisons difficult.

As described above, different methods have been used to evaluate trend forecasts depending on the purpose. In particular, methods using proxy indicators often use domain-specific definitions, making it difficult to conduct a consistent evaluation of diverse domains. For example, while product demand can be used as a proxy indicator in the clothing domain, it cannot be measured in the personal domain. Moreover, a trend forecast does not necessarily represent a time-series forecast. In other words, no consistent evaluation method has been established to predict trending entities. Addressing these limitations, this study aims to construct an evaluation dataset that can be used to evaluate various trend forecasting methods.

Table 2 presents a comparative analysis between the datasets used in previous studies and Trend Dataset constructed in this research, with respect to the six perspectives required for the protocol of evaluating trend forecasting methods. These six perspectives encompass the four issues for the evaluation protocol of trend forecasting methods and the the three essential attributes of that protocol, all of which were presented in § 1. Table 2 indicates that Trend Dataset adequately addresses all six perspectives, making it a unique dataset.

3 Proposed Method

The following section describes the methodology for constructing a dataset used to evaluate the trend forecasting methods. We collect entities with varying likelihoods of trending, and measure recognition rate time series based on a crowdsourcing questionnaire survey and Internet search volume. Based on the obtained time series, we annotate three attributes: trending status (trending or non-trending), degree of trending, and trend period. The proposed method consists of three phases: setup of an entity master, measurement of recognition rate time series, and annotation of trend attributes. Figure 1 presents an overview of the proposed method.

3.1 Setup of Entity Master

We first construct an entity master, a collection of the entities comprising the dataset, with Wikipedia articles serving as entities. The objective is to collect new entities, both trending and non-trending, from a wide range of domains without significant imbalance. As the random choice of articles incurs a bias towards popular domains and non-trending entities, we propose a method to identify interest patterns and domains utilizing the article metadata. An interest pattern represents the changes in interest regarding an article among Wikipedia users. In other words, interest patterns can be utilized to gauge the likelihood that an entity has been trended. Entities are chosen uniformly from each domain and interest pattern to achieve a balanced collection.

The classification of domains and interest patterns is presented in Tables 3 and 4. We define 21 domains[5], with each combining a category and subcategory. We assign a domain for each article using the similarity between articles and keywords unique to each domain. We define three classes for interest patterns: `positive`, `negative`, and `negative-popular`. The assignments to `positive` and `negative` categories are executed systematically using the

[5]We define domain classification with reference to Sekine's extended named entity hierarchy (Sekine, Sudo, and Nobata 2002).

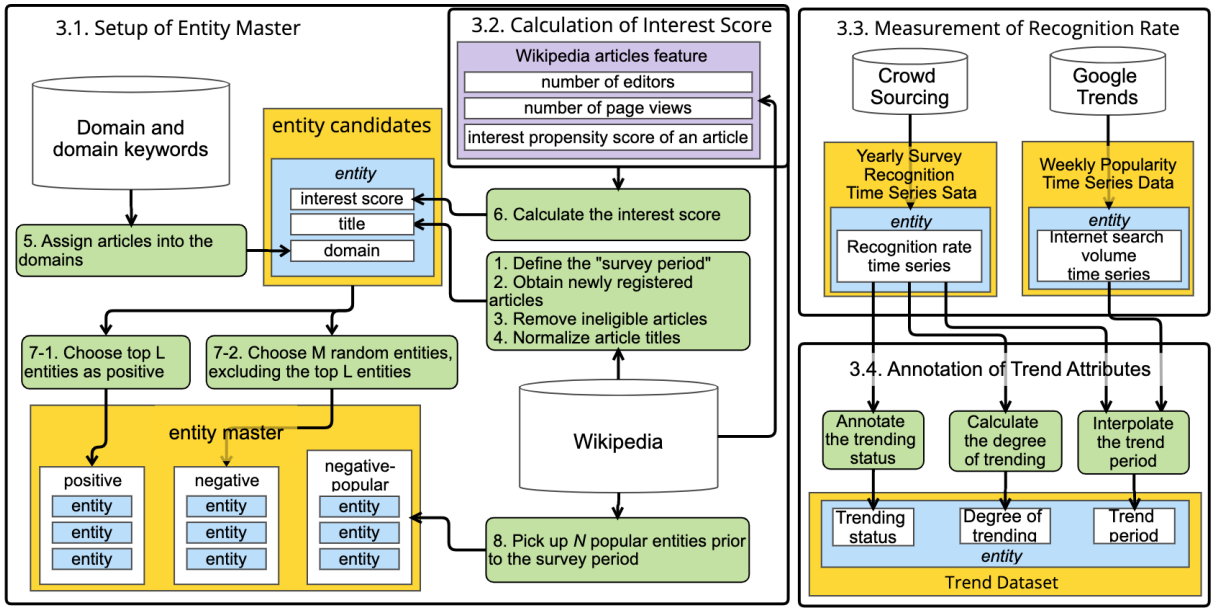

Figure 1: Overview of the proposed method.

Category	Subcategory
Location/ Geography	City/region/landmark
Organization	Restaurant/facility, company/brand
Person/ Group	Politician/political party, researcher, athlete, actor/actress, celebrity/entertainer/comedian, music band/music group
Product	Cosmetics, daily necessities, clothing, beverage, foodstuff, others
Art/Content	Game, publication, comic/animation, movie, broadcast program, music

Table 3: List of entity domain categories.

Interest pattern	Definition
`positive`	Rapidly attracted interest within the survey period
`negative`	Did not attract interest
`negative-popular`	Widely known before the survey period

Table 4: List of entity interest pattern classes.

interest score. Interest score (§ 3.2) is calculated from the number of editors, number of views, and the interest propensity score which is measured using the article texts. The entities assigned to the `negative-popular` class is manually chosen with reference to the number of page views of the article[6].

The purpose of adding `negative-popular` entities to the master is to increase the difficulty of the trend forecasting task. Specifically, we want to ensure that simplistic trending status prediction method, which solely uses the temporal average of the recognition rate (or its proxy indicator), should be ineffective.

The procedure for constructing the entity master is as follows:

1. Define the period we want to cover for the trend forecasting task, hereafter referred to as the "survey period".
2. Obtain newly registered Wikipedia articles during the survey period.
3. Remove articles that can not be regarded as entities. Specifically, we exclude articles that are intended for redirects, listings, or disambiguation, and those that include the year in the article title.
4. Normalize article titles by removing any parentheses at the ends. If articles with duplicate normalized titles are present, we retain the article with the earliest registration date.
5. Assign articles into the domain with the highest similarity. The similarity between a domain and an article is measured using the manually defined domain-specific keywords[7] and the first sentence of the article. Specifically, we convert keywords and the first sentence into distributed representations and compute cosine similarity[8].

[6]A systematic choice of most-viewed articles is not recommended because widely recognized entities (e.g., vacuum cleaners, Nintendo) do not necessarily have top page views.

[7]Words that frequently appear in the first sentence of articles belonging to the domain are selected as keywords. For example, the keywords of the "People–Politician/Political Party" domain are {politician, political party, minister, prime minister, president, legislator}.

[8]We use Word2Vec (Mikolov et al. 2013) for the keywords, and Smoothed Inverse Frequency (Arora, Liang, and Ma 2017) for the first sentence of the article, respectively.

Type of entity	Correct response
Products and organizations that are widely adopted in society	I knew about this since before the survey period
Incidents and disasters that occurred during the survey period and were widely reported	I knew about this since the year the incident/disaster occurred
Names of extremely low-profile persons or places	I don't know

Table 5: Entity types and expected correct responses used for validation questions.

Item	Setting
Format	Multiple choice
Question	When did you learn about `entity name`? If you do not know `entity`, select "I don't know".
Choices	I don't know, I knew since before the survey period, I knew since year t[4]

Table 6: Survey questions used for recognition survey.

6. Calculate the interest score (§ 3.2) using the article's number of editors, number of views, and interest propensity score. A higher score indicates a greater likelihood of a surge in interest among Wikipedia users.
7. Collect top L entities with the highest interest scores for each domain, which correspond to entities with a `positive` interest pattern. Similarly, choose M entities randomly from the remaining entities, which correspond to entities with `negative` interest pattern.
8. For each domain, manually pick N entities which has been popular since before the survey period. These correspond to the entities with `negative-popular` interest pattern.

The set of entities collected through the above procedure is called the entity master. The entity master is expected to contain entities from various domains and degrees of interest.

3.2 Calculation of Interest Score

The interest score is an indicator that suggests the degree of a surge in interest for Wikipedia articles. Specifically, it is calculated using the number of editors, page views, and interest propensity score of an article as features, defined as the distance from the hyperplane through the minimum of each feature. Let x_{e} be the logarithm of the number of unique editors, and x_{v} be the logarithm of the daily average number of views during the survey period. The interest propensity score x_{t} is the cosine similarity of the articles and words expressing trends[9], each converted to their respective distributed representations[10].

Let $\mathbf{x}_{\mathrm{wiki}}(a_i)$ be a three-dimensional vector with the aforementioned features for article a_i, and the plane $\mathbf{w}$ passing through the minimum values $x_{\mathrm{e},0}, x_{\mathrm{v},0}, x_{\mathrm{t},0}$ of each be given by

$$\mathbf{w} = \left(\frac{1}{x_{\mathrm{e},0}}, \frac{1}{x_{\mathrm{v},0}}, \frac{1}{x_{\mathrm{t},0}}\right)^T.$$

Then the interest score $s(a_i)$ is defined as the signed distance from the plane

$$s(a_i) = \frac{\mathbf{w}^T \mathbf{x}_{wiki}(a_i) - 1}{||\mathbf{w}||}. \quad (1)$$

3.3 Measurement of Recognition Rate Time Series

The recognition rate time series is measured for each entity in the entity master (§ 3.1). Specifically, a crowdsourced questionnaire survey is used to measure public recognition rates on an annual basis. Subsequently, the weekly recognition rate is interpolated using Internet search volume.

Measurement of public recognition rate We use a crowdsourcing questionnaire survey and measure the public awareness of entities in conjunction with the point in time they are recognized. Table 6 shows the crowdsourcing survey questions.

To ensure data quality, we exclude unreliable crowd workers. Specifically, we use the verification questions and the inspection of response patterns to identify such workers. Verification questions are created using very high- or low-profile entities, for which the recognition and corresponding year should be obvious to the general public. Table 5 shows the entity types and gold standards used as verification questions. The response pattern inspection is conducted to examine whether the aggregated responses are statistically deviant. Specifically, a worker is deemed to be unreliable if they consistently respond with "I don't know" or "I knew since before the survey period" for all questions, as such a response pattern is implausible unless there is a significant imbalance of the entities. We refer to the remaining workers as valid workers.

Upon completion of the questionnaire survey, we compute the yearly recognition rates. The response of which year the crowd workers knew the entity allows us to measure the percentage of crowd workers who recognize the entity by the end of each year. In essence, the questionnaire-based recognition rate $C_q(t)$ at each end-of-year time t within the survey period is calculated as

$$C_q(t) = \frac{W(t)}{W_{whole}}, \quad (2)$$

where $W(t)$ is the total number of workers who responded "I knew since before year t"[11], and W_{whole} is the number of

[9] We use synonyms for "trend", such as "boom" and "fashion".

[10] We use Wikipedia2Vec (Yamada et al. 2016). Wikipedia2Vec is a distributed representation model that embeds articles and words in the same space.

[11] For example, let the survey period be from 2015 to 2018. In this setting, $W(t = 2016)$ is the sum of "I knew since before 2015" and "I knew since 2016".

valid workers.

Interpolation by Internet search volume The crowdsourced questionnaire-based recognition rate (public recognition rate) has an annual basis, which is not a sufficient resolution to annotate the trend period. Therefore, we propose a method to interpolate the weekly recognition rate by assuming that the cumulative sum of Internet search volume according to Google Trends is proportional to the recognition rate. Note that internet search volume does not affect the annotation of trending status (assigned by the change in questionnaire-based recognition rate; see § 3.4); it affects the annotation of trend period. The rationale for restricting the utilization of Internet search volume for interpolation is that Google Trends may not necessarily reflect the public recognition rate, as we explained in the introduction (§ 1).

First, we obtain the weekly Google Trends data, with the query parameters shown in Table 7. The query string is the entity name. Then, Google Trends-based recognition rate C_{gt} at time t is defined as the cumulative sum of the Google Trends values P_k until each point in time:

$$C_{gt}(t) = \frac{\sum_{k=1}^{N_t} P_k}{\sum_{k=1}^{N_{whole}} P_k}, \tag{3}$$

where N_{whole} is the number of time series data points and N_t is the number of data points up to time t. Next, We adjust the Google Trends-based recognition rate $C_{gt}(t)$ in order to align with the questionnaire-based recognition rate $C_q(t)$. Specifically, the time series between the start of the year T_i and the end of the year T_{i+1}, included in the survey period, are linearly transformed to match $C_q(t)$. Let $C_c(t)$ be the interpolated recognition rate. The interpolation formula is defined as follows:

$$\alpha = \frac{C_q(T_{i+1}) - C_q(T_i)}{C_{gt}(T_{i+1}) - C_{gt}(T_i)}, \tag{4}$$

$$C_c(t) = C_q(T_i) + \alpha \left(C_{gt}(t) - C_{gt}(T_i)\right); \quad T_i \leq t < T_{i+1}. \tag{5}$$

3.4 Annotation of Trend Attributes

The following section describes the method for annotating trend attributes based on the measured recognition rate time series. We annotate the three trend attributes: trending status, degree of trending, and trend period.

Trending status The questionnaire-based recognition rate (§ 3.3) is used to annotate the trending status of each entity. Specifically, an entity is annotated as trending if 1) the difference of the questionnaire-based recognition rate at the end of the survey (t_{end}) and that at the start of the survey (t_{start}) is greater than the threshold τ, and 2) the questionnaire recognition rate at the start of the survey is smaller than the threshold τ_0:

$$C_q(t_{end}) - C_q(t_{start}) \geq \tau \wedge C_q(t_{start}) < \tau_0. \tag{6}$$

Thus, an annotation of "trending" means that the entity has been low profile before the survey period but rapidly gained recognition during the survey period.

Parameter	Value
Query	Entity name
Query type	Topic[12]
Period	*Same as survey period*
Interval	Weekly
Category	All categories

Table 7: Query parameters of Google Trends.

Degree of trending The questionnaire-based recognition rate (§ 3.3) is used to calculate the degree of trending for each entity. Specifically, the degree of trending p is the difference between the questionnaire-based recognition rate at the end of the survey period t_{end} and the start of the survey period t_{start}:

$$p = C_q(t_{end}) - C_q(t_{start}). \tag{7}$$

Trend period Specific to entities that are annotated as "trending", we annotate the trend period using the interpolated recognition rate (§ 3.3). Specifically, the trend period is defined as a percentile of the interpolated recognition rate $C_c(t)$. That is, we define the dates $t_s = Q_{q_s}(C_c)$ and $t_e = Q_{q_e}(C_c)$ corresponding to the percentiles q_s and q_e ($0 < q_s < q_e < 1$) as the trend period $[t_s, t_e]$. Note that $t_s \leq t_e$ always holds true because, by definition, the interpolated recognition rate $C_c(t)$ increases monotonically with respect to t.

4 Construction of Trend Dataset

In this section, we report the results of constructing a dataset (hereafter known as "Trend Dataset") for Japan using the proposed method described in § 3.

4.1 Setup of Entity Master

We defined survey period from January 1, 2015, to April 30, 2019. We extracted the new articles from the Japanese version of Wikipedia (as of April 20, 2019). The number of page views for articles was collected using the Wikimedia REST API[13]. HottoSNS–w2v (Matsuno, Mizuki, and Sakaki 2019) was adopted for the Word2Vec model used for domain category assignment. The Wikipedia2Vec model, used for calculating interest propensity scores (§ 3.2), was trained using the aforementioned Wikipedia dump. Entities with `negative-popular` interest patterns were collected by the dataset creators, among whom were the authors of this study. The number of candidate entities was $1,291$. Then, we collected $L = 10$–11, $M = 5$, and $N = 2$–5 `positive`, `negative`, and `negative-popular` entities, respectively, from each domain, for a total of 400 entities as the entity master.

4.2 Measurement of Recognition Rate Time Series

Public recognition rate For all entities in the entity master, we surveyed the questionnaire-based recognition rate in

[13] https://wikimedia.org/api/rest_v1/

Item	Settings
# of entities	400
# of responses per entity	510
# of responses per worker	20—200
Question	*Refer to Table 6*
Choices	I don't know, I knew since before 2015, I knew since 2016, I knew since 2017, I knew since 2018, I knew since 2019

Table 8: Crowdsourcing task design for recognition survey.

Gender	Age	# of workers	# of valid workers
Male	15–29	254	201
Female	15–29	255	210
Male	30–39	253	244
Female	30–39	275	234
Male	40–59	246	222
Female	40–59	257	240
Total		1,540	1,351

Table 9: Number of crowd workers by age and gender.

the Japanese population using Yahoo! Crowdsourcing[14]. Table 8 shows the task design. Crowd workers were recruited uniformly from each gender and age group to diversify the demographics of the questionnaire survey. Workers who answered the verification questions incorrectly were excluded during the survey, whereas those who exhibited abnormal response patterns were excluded after the survey was completed. Table 9 presents the number of workers, indicating a valid worker rate of $1351 \div 1540 = 87.7\%$.

The questionnaire-based recognition rate yearly time series $C_q(t)$ (Eq. 2) was calculated using the responses given by the crowd workers. Note that the survey were conducted in mid-September 2019; thus, the choice "I knew since 2019" was aligned to August 31, 2019, not December 31, 2019. To assess the reliability of the crowd workers' responses, we examined the distribution of the percentage of entities for which the response was "I knew" for each crowd worker, with results shown in Figure 2. The distribution of "I knew" response percentages were expected to be in the 20–75% range (mean 47%) if most workers responded in good faith. This assumption stems from the ratios of interest rate classes; we expect that, on average, workers will respond "I knew" for the `negative-popular` entities and "I don't know" for the `negative` entities. The former and latter comprise approximately 20% and 25% of all entities (Table 10). Figure 2 indicates that the distribution adheres, on overall, to the assumption for both genders and all age groups[15]. Consequently, we posit that the majority of crowd workers answered in good faith.

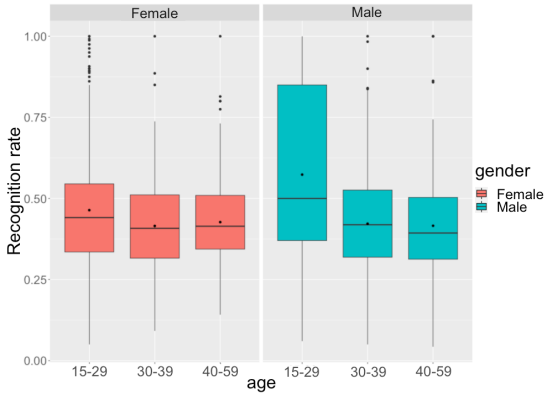

Figure 2: Distribution of percentage of entities that each crowd worker answered "I knew".

Interpolation by Internet search volume We used Google Trends to get the Internet search volume. Japan was specified as the region of interest, with the data collection time period set from September 28, 2014 to August 25, 2019. Of the 400 entities, data were obtained for all entities annotated as trending (80, see Table 10). Consequently, we could annotate the trend period for trending entities as expected. Data were obtained for 288 out of the 320 non-trending entities.

Interest pattern	Trending	Non-trending	Total
`negative`	0	105	105
`negative-popular`	0	84	84
`positive`	80	131	211
Total	80	320	400

Table 10: Number of entities by the interest pattern and the trending status.

4.3 Annotation of Trend Attributes

Using the measured recognition rate time series, we annotated the three trend attributes (§ 3.4): trending status, degree of trending, and trend period. The thresholds[16] for determining trending status were configured to $\tau = 0.25$ and $\tau_0 = 0.3$. The percentiles defining the trend period were configured to $q_s = 0.25$ and $q_e = 0.75$. As shown in Table 10, we annotated 80 entities as trending and 320 entities as non-trending, respectively. Table 1 (in §1) shows examples of entities annotated with trend attributes. Figures 3 and 4 show the distributions of the start date, end date, and length of the trend period, respectively. Overall, the trend

[14] As this service is for the domestic market, the majority of crowd workers are considered to be Japanese. https://crowdsourcing.yahoo.co.jp/

[15] Approximately 70 male workers aged 15–29 responded "I knew" more than 80% of all questions. As they exceeded the upper limit of reasonable assumption, a dishonest response is suspected in these cases. However, as they accounted for only about 5% of all workers, we decided not not exclude them.

[16] We utilized the rule-of-thumb in the marketing industry, known as "Chasm," to configure the thresholds. As an example, the values of $\tau = 0.25$ and $\tau_0 = 0.3$ roughly correspond to crossing the chasm and achieving wide recognition among the early majority.

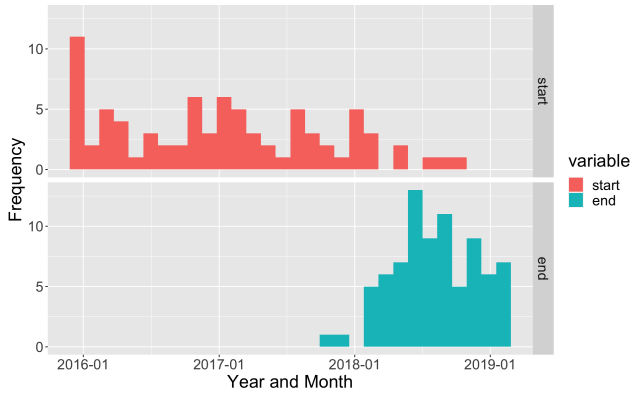

Figure 3: Histogram of trend period start and end.

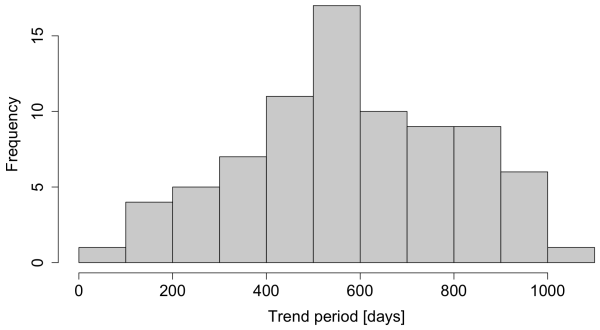

Figure 4: Histogram of trend period lengths.

start dates were uniformly distributed throughout the survey period. The distribution of the trend period length exhibited a bell curve within the range of 100–1000 days, with the most frequently observed length being 500–600 days.

5 Analysis

In this section, we validate and examine the characteristics of the Trend Dataset using statistical analysis. Specifically, we investigate whether the interest patterns assigned using Wikipedia metadata are consistent with public recognition rates measured using crowdsourcing. We also investigate the differences in the degree of trending across domains and trend status. Finally, we examine whether an increase in Internet search volume measured using Google Trends is consistent with an increase in public recognition rate.

5.1 Consistency between Interest Patterns and Public Recognition Rates

Table 10 shows the number of entities for each combination of interest pattern and trending status. As seen in Table 10, all trending entities exhibit a `positive` interest pattern, suggesting that assigning interest patterns a priori contributed to the efficient collection of trending entities. Figure 5 shows the distribution of the degree of trending and recognition rates at the end of the survey. It indicates that the degree of

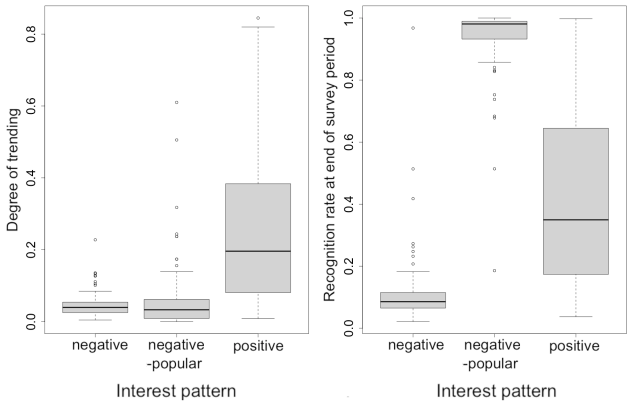

Figure 5: Distribution of degree of trending (left) and the recognition rate at the end of survey period (right) by interest patterns.

trending, i.e., the increase in recognition rate during the survey period, is greater for the `positive` entities and lower for the `negative` and `negative-popular` entities. In addition, we can observe that the recognition rate of the `negative` entities remains low, whereas that of the `negative-popular` entities has been high since before the survey period. These observations are consistent with the characteristics expected for each class of interest pattern. We also tested whether the difference of trending degree between the `positive` group and `negative` $\bigcup$ `negative-popular` groups was statistically significant (Mann—Whitney U test (two-sided)). We confirmed that the degree of trending within the `positive` group is significantly higher[17] ($U = 19112, p =$ 2.2e-16). These results indicate that the estimated changes in interests regarding a specific article among Wikipedia users are highly consistent with the changes in public recognition rate.

5.2 Trending Level

Figure 6 (left) shows a distribution of the degree of trending in relation to trending status, confirming that the degree of trending of trending entities is higher than that of non-trending entities. Figure 7 shows the distribution of the degree of trending by domain and trending status, demonstrating that the phenomenon applies to all domains. However, the level of trending degree varies among domains. Specifically, the degree of trending is low for the "Person/Group–Researcher", "Location/Geography–City/Region/Landmark", and especially "Product" categories. In fact, no entities were annotated as trending in these domains, suggesting that some domains are less likely to be of interest to the public and have an inherent upper limit to gaining recognition. We defer the development of a suitable annotation methodology for entities in niche domains to future work. Potential approach to this end involve introducing the domain-specific thresholds (Eq. 6) for de-

[17] In the case of the `negative` and `negative-popular` groups, significant differences in degree of trending was rejected at the 5% level ($U = 5083, p = 0.072$)

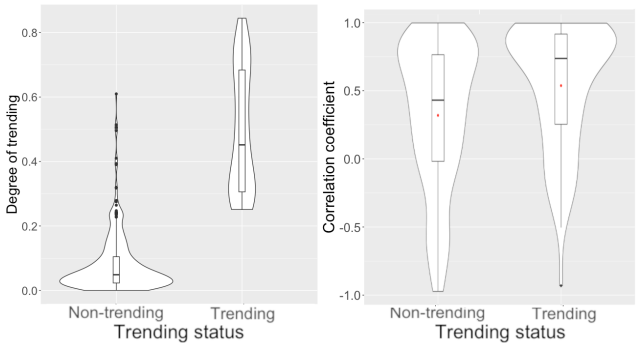

Figure 6: Distribution of degree of trending by trending status (left) and the distribution of the correlation coefficients by trending status (right). Those violin plots shows the distribution of degree of trending or correlation coefficients per entity, boxes represent the quartiles, and horizontal lines and dots in the box figure represent the median and mean, respectively.

termining the trending status. Another promising approach involves asking crowd workers to respond to their interest in the domain, as well as their recognition of the entity.

5.3 Consistency between Google Trends and Public Recognition Rate

Interpolating the weekly recognition rate using Google Trends stems from the assumption that the cumulative sum of Google Trends is proportional to the public recognition rate. We evaluated the validity of this assumption. Specifically, we analyzed Pearson's correlation coefficient between the first difference of the questionnaire-based recognition rate yearly time series: $C_q(T_{i+1}) - C_q(T_i)$ (Eq.2) and the first difference of the Google Trends-based recognition rate yearly time series: $C_{gt}(T_{i+1}) - C_{gt}(T_i)$ (Eq. 3). We divided entities subject to analysis into two groups: trending (80 entities) and non-trending for which Google Trends data was obtained (288 entities). The rationale for dividing them by trending status is that the Google Trends-based recognition rate only affects annotations of trend periods of trending entities. Figure 6 (right) and Table 12 show the distribution and summary statistics of the correlation coefficients, respectively. Among trending entities, the mode and median of the correlation coefficients were approximately 0.9 and 0.74, respectively. It indicates that the assumption regarding the proportionality between Google Trends and the public recognition rate is generally valid for trending entities. Therefore, the proposed annotation method for the trend period is considered reasonable. The correlation coefficients for the non-trending entities are relatively low and more dispersed compared to those for the trending entities, suggesting that Internet search volume does not necessarily reflect the public recognition rate for entities that do not experience a surge in public recognition.

6 Discussion

We discuss the proposed method's effectiveness and limitations based on the analysis of Trend Dataset. We also propose evaluation tasks for trend forecasting methods using the dataset.

6.1 Effectiveness and Limitations

One of the strengths of the proposed method is its adaptability. Although we constructed Trend Dataset for Japanese society, the proposed method applies to other societies or countries. Strictly speaking, proposed method can be applied in countries with sufficient country-specific Wikipedia articles and Google Trends data. Although we have used a crowdsourcing to measure public recognition, alternative sample surveys can also be employed. The proposed method has several limitations. The first limitation is the duration of the survey period. As the year of recognition is asked retrospectively in the sample survey, longer survey periods may reduce the reliability of the responses. The second limitation is the comprehensiveness of domains (§ 5.2). It is necessary to ensure appropriate sample survey demographics, questions, and criteria to determine trending status within domains of low public interest. The third limitation is the accuracy of the trend period. Although Google Trends and recognition survey results are consistent for most trending entities (§ 5.3), there are certain exceptions. The annotation of the trend period by weekly resolution leaves room for further improvement in accuracy.

6.2 Trend Forecasting Method Evaluation Tasks

We now present several tasks for evaluating trend forecasting methods using Trend Dataset. Specifically, three evaluation tasks can be performed: binary classification of trending status, ranking of degree of trending, and prediction of the trend period. Table 11 lists the trend attributes used for each task. Binary classification of trending status is the task of predicting whether an entity will be trending during a predetermined time frame. This task is particularly useful when we have a list of interested entities in advance; for example, predicting the popularity of a newly released product. A ranking task is a task that predicts the degree of trending of each entity over a specific time frame and ranks the entities in descending order. This task is suited for comparing the trending levels among products or services; for example, predicting the top five fashions that will be popular in next winter. The trend period prediction task is a task that extrapolates the recognition rate time series and predicts the start and end dates of the trending spans. This task is suited when we know the trending entities in advance; for example, predicting how long current blockbuster movies stay in the box office.

6.3 FAIRness Data Availability Statement

The provided dataset (Trend Dataset) conforms to the FAIR principles, can be searched and accessed via Zenodo (DOI: https://doi.org/10.5281/zenodo.7014424), and is licensed under CC BY 4.0. In addition, the dataset is shared

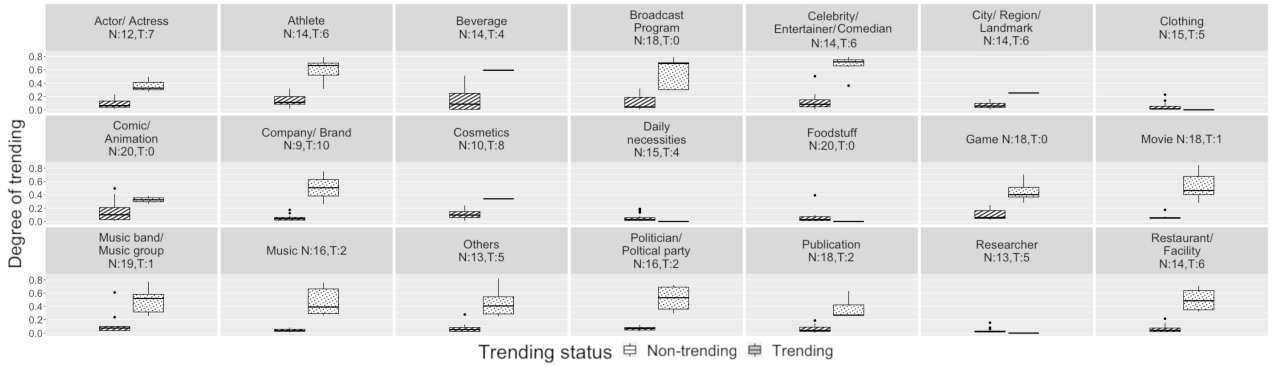

Figure 7: Distribution of the degree of trending by trending status for each domain. "N" and "T" in the title of each facet indicates the number of non-trending and trending entities, respectively. Additionally, trending status is represented by stripe and dot hatch.

Task objective	Task type	Trend attributes		
		Trending status	Degree of trending	Trend period
Which entity will trend	Binary classification	√		
How much it will trend	Ranking	√ (optional)	√	
When it will trend	Time series prediction	√		√

Table 11: Evaluation tasks that use Trend Dataset.

Trending status	max	min	mean	median
Trending	1.00	-0.93	0.54	0.74
Non-trending	1.00	-0.97	0.32	0.43

Table 12: Summary statistics of the correlation coefficients between Google Trends-based and questionnaire-based recognition rates by trending status.

as a set of csv files with a summary describing its configuration and published in a reusable and interoperable format. Church et al. suggested that one of the factors supporting the rapid progress in applied machine learning research fields, such as NLP and CV, is the standardization of tasks and public datasets, which allows proposed methods to be evaluated in a quantitatively comparable form (Church and Hestness 2019). In trend forecasting studies, such attempts have been limited to specific domains, such as finance, with no attempts made to target diverse domains. We hope this study will encourage steady progress in the social science research field of trend forecasting.

7 Conclusion

In this study, we tackled four limitations in the evaluation methods of conventional trend forecasting: usage of both trending and non-trending entities, trend period prediction, standardization of domains and entities subject to evaluation, and adoption of a highly reliable public recognition rate indicator. After describing a desired property for evaluating trend forecasting methods, we proposed a method to construct a dataset that can be used for such an evaluation. There are four main features of the proposed method. First, a questionnaire-based recognition rate is used as the gold standard for annotating trending status. Second, a collection of entities from a wide range of domains while covering both trending and non-trending, without significant imbalance. Third, an efficient dataset collection and annotation process. Fourth, trend period annotation on a weekly resolution through interpolation using Internet search volume data. We conducted the proposed method and compiled Trend Dataset. We then analyzed its quality and characteristics to assess the validity of the proposed method. Specifically, we confirmed that changes in the interests of entities estimated using Wikipedia article metadata are consistent with the public recognition rate measured by a questionnaire, enabling the efficient collection of trending entities a priori. Also, we confirmed a good correlation between Internet search volumes and public recognition rates for trending entities, enabling annotation of trend periods on a weekly resolution. Consequently, we demonstrated that we could annotate trending status, degree of trend, and trend period without relying on a large-scale and frequent social survey. We are aware of several limitations, including the length of the survey period, domain comprehensiveness, and annotation accuracy of the trend period. Despite these limitations, the Trend Dataset enables the standardized evaluation for trend forecasting methods: predicting which entity will trend, how much it will trend, and when a trend will occur. We expect the dataset and dataset construction procedure will promote further progress in the social science research field of trend forecasting.

Ethical Statement

The authors have carefully read and adhered to the AAAI CODE and ICWSM Guidelines with respect to this study and the accompanying dataset. From the perspective of pri-

vacy protection, this study assumes that personally identifiable information online is processed in an anonymized format. For example, editor information included as metadata on Wikipedia is processed anonymously. In this case, the object of anonymization is not limited to real names but includes Web service identifiers that can uniquely identify web service users. Furthermore, as described in § 3.3, we ensured the reliability of all responses through screening and used ethical considerations in the crowdsourcing questionnaire survey of this study. Specifically, we promised to guarantee the voluntary nature of responses, the right to withdraw during the response process, and data anonymity; in addition, we provided a full explanation of the task to each crowd worker before conducting a questionnaire. All other information in the dataset was obtained from publicly available data on the Internet.

References

Arora, S.; Liang, Y.; and Ma, T. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *Proceedings of 5th International Conference on Learning Representations*.

Bandari, R.; Asur, S.; and Huberman, B. 2012. The pulse of news in social media: Forecasting popularity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, 26–33.

Chambers, J. C.; Mullick, S. K.; and Smith, D. D. 1971. How to choose the right forecasting technique. *Harvard Business Review*, 49: 45–71.

Cheng, J.; Adamic, L. A.; Kleinberg, J. M.; and Leskovec, J. 2016. Do Cascades Recur? In *Proceedings of the 25th International Conference on World Wide Web*, 671–681.

Choi, H.; and Varian, H. 2012. Predicting the Present with Google Trends. *Economic Record*, 88(s1): 2–9.

Church, K. W.; and Hestness, J. 2019. A survey of 25 years of evaluation. *Natural Language Engineering*, 25(6): 753–767.

Dogan, A.; and Birant, D. 2021. Machine learning and data mining in manufacturing. *Expert Systems with Applications*, 166: 114060.

Forslund, H.; and Jonsson, P. 2007. The impact of forecast information quality on supply chain performance. *International Journal of Operations & Production Management*, 27(1): 90–107.

Jang, B.; Kim, I.; and Kim, J. W. 2021. Long-Term Influenza Outbreak Forecast Using Time-Precedence Correlation of Web Data. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Li, X.; Xie, Q.; Huang, L.; and Yuan, Z. 2017. Twitter data mining for the social awareness of emerging technologies. In *Proceedings of Portland International Conference on Management of Engineering and Technology*, 1–10. IEEE.

Matsuno, S.; Mizuki, S.; and Sakaki, T. 2019. Constructing of the word embedding model by Japanese large scale SNS + Web corpus. *Proceedings of the Annual Conference of JSAI*, JSAI2019: 4Rin113–4Rin113.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, 3111–3119.

Murray, P. W.; Agard, B.; and Barajas, M. A. 2018. Forecast of individual customer's demand from a large and noisy dataset. *Computers & Industrial Engineering*, 118: 33–43.

Pournader, M.; Ghaderi, H.; Hassanzadegan, A.; and Fahimnia, B. 2021. Artificial intelligence applications in supply chain management. *International Journal of Production Economics*, 241: 108250.

Rousidis, D.; Koukaras, P.; and Tjortjis, C. 2020. Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9): 6279–6311.

Roy, S. D.; Lotan, G.; and Zeng, W. 2015. The attention automaton: Sensing collective user interests in social network communities. *IEEE Transactions on Network Science and Engineering*, 2(1): 40–52.

Sayyadi, H.; Hurst, M.; and Maykov, A. 2009. Event Detection and Tracking in Social Streams. *Proceedings of the International AAAI Conference on Web and Social Media*, 3(1): 311–314.

Sekine, S.; Sudo, K.; and Nobata, C. 2002. Extended Named Entity Hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA).

Skenderi, G.; Joppi, C.; Denitto, M.; and Cristani, M. 2021. Well Googled is Half Done: Multimodal Forecasting of New Fashion Product Sales with Image-based Google Trends. *CoRR*, abs/2109.09824.

Tsur, O.; and Rappoport, A. 2012. What's in a hashtag? Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 643–652.

Yamada, I.; Shindo, H.; Takeda, H.; and Takefuji, Y. 2016. Joint Learning of the Embedding of Words and Entities for Named Entity Disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 250–259.

Yu, L. L.; Asur, S.; and Huberman, B. A. 2015. Trend Dynamics and Attention in Chinese Social Media. *American Behavioral Scientist*, 59(9): 1142–1156.

Yu, S.; and Kak, S. C. 2012. A Survey of Prediction Using Social Media. *CoRR*, abs/1203.1647.