Emerging Product Topics Prediction in Social Media without Social Structure Information

Conference Paper · April 2018

DOI: 10.1145/3184558.3191625

Man-Kwan Shan
National Chengchi University
83 PUBLICATIONS **1,230** CITATIONS

Emerging Product Topics Prediction in Social Media without Social Structure Information

Sinya Peng
National Chiao Tung University
Taiwan
sinyaya99.cs04g@nctu.edu.tw

Vincent S. Tseng
National Chiao Tung University
Taiwan
vtseng@cs.nctu.edu.tw

Che-Wei Liang
Industrial Technology Research Institute
Taiwan
jared@itri.org.tw

Man-Kwan Shan
National Chengchi University
Taiwan
mkshan@nccu.edu.tw

ABSTRACT

Social media provides a vast continuous supply of dynamic and diverse information contents from the crowd, which serves as useful resources for predictive analytical applications. Although there exist already a number of studies on emerging topics detection, they focused on modelling of textual contents and emerging detection mechanism over topic popularity. To meet the real-life demands, prediction of emerging product topic, rather than detection, in the early stage is required. Besides, despite that some relevant studies considered social structure information, they suffer from the assumption that the complete network is available and the diffusion process only depends on social influence among members of networks. Moreover, not all social media sites provide the functionality to facilitate the development of online social networks. In this paper, we tackle the problem of emerging product topics prediction in social network with implicit networks. Two tasks, one for long-term forecast in pre-production stage and the other for short-term forecast in post-release stage, are investigated. We present a novel framework named Emerging Topics Predictor (ETP). Two novel features, namely author diversity and competition features, are also proposed to accommodate the diffusion process with implicit networks based on the rationale of product marketing. Through empirical evaluation on movie reviews from two real social media sites, ETP is shown to provide effective and efficient performance in predicting the emerging topics as early as possible. In particular, the experiment results show the promising effect of author diversity in emerging prediction. To the best of our knowledge, this work is among the very first studies on emerging product topic prediction in social media with considerations of implicit networks.

CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; *Social networks*; Business intelligence • **Computing methodologies** → Information extraction

KEYWORDS

emerging product predict; implicit network

WWW '18 Companion April 23-27, 2018, Lyon, France.
© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.
ACM ISBN 978-1-4503-5640-4/18/04.
DOI: https://doi.org/10.1145/3178876.673

1 INTRODUCTION

With the rapid growth of social media sites, large amounts of customer reviews on products are authored, posted, spread and propagated over online social networks. The instantaneity, accessibility, and popularity of social media provide the opportunity for collection of customer opinions, estimation of product sales, and prediction of products' future trends by discovering patterns from customer reviews.

An increasing number of studies have investigated the dynamics of online word-of-mouth. A positive relationship between online customer reviews and product sales was found [10]. In parallel, other researches also show that the volume of reviews has a significant effect on new product sales in the early period and such effect decreases over time [16].

Numerous recent studies have been devoted to tackling the problem on detection of emerging topics from social media data, specifically, opinions/items that have the potential for high popularity due to increased public interests while they are not popular currently. Existing researches on emerging topic detection focused on the dynamic modeling of textual contents and the emerging detection mechanism over topic popularity. These works are constrained in the detection of emerging topics. To meet the realistic demand of providing useful insight by predicting emerging product topics in the early stage or before they are materialized, prediction (rather than detection) of emerging product topics as early as possible is needed.

Some other research has devoted to the prediction of spike time of information cascades over social media. An information cascade occurs as information propagates through friends. Most research reported that temporal and social structure features are key indicators for prediction [3, 8, 11]. These problems have been solved in the aspect of spike prediction on information cascades, but remain widely open on the emerging topic prediction with implicit network which is shortfall of network structural properties. Furthermore, not all social media sites provide the functionality to facilitate the development of online social networks. The social network over which diffusion takes place is unknown. These social news aggregation and discussion sites without explicit social network have tremendous customer reviews on products. One example is Reddit, which had 542 million monthly visitors, ranking as the 4th most visited website in U.S. and the 8th in the world, as of 2017. Reddit is composed of thousands of user created and moderated subreddits, which are discussion boards covering a variety of topics including news, movies, music, science, and so on.

Based on the observations as above, in this paper, we tackle the problem of emerging product topic prediction in social network with implicit networks. Here, an emerging product topic indicates a topic of products that gathers a significantly growing number of reviews over social media (more detailed definition will be given in Section III). Two prediction tasks are investigated as follows:

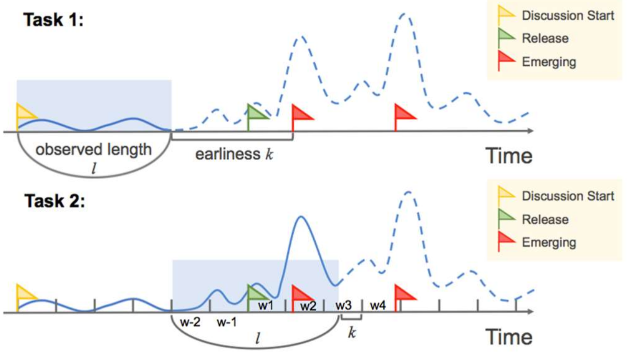

Figure 1: The Illustration of Two Prediction Tasks.

- **Task 1:** Given a set of product review streams in social media site, which product topics will become emerging in the future? How can we predict it accurately and as early as possible?
- **Task 2:** Given a set of product review streams in social media site, which product topics will become emerging in the next predicting time interval? How can we predict it accurately and as early as possible?

Task 1 is designed for long-term prediction during the early stage of product development. For example, the production of a movie is an expensive, risky endeavor. While movie fans tend to paid attention to relevant news early in the pre-production stage, pre-production analysis and prediction from movie discussions authored by fans are helpful for decision makers of the film industry. In contrast, Task 2 focuses primarily upon short-term prediction in the stages of product life cycle. For example, in the third week after a movie is released, along with the information of simultaneous released movies, the decision maker would like to make the prediction for the fourth week.

To tackle the problem of these two tasks over social media with implicit network, a novel framework named *Emerging Topics Predictor* (*ETP*) is proposed. In particular, we proposed a new type of features, author diversity, to deal with the diffusion process with implicit network. Moreover, Products, which perform the same function, compete against each other. However, to the best of our knowledge, no work has incorporated the competition features into the prediction model for emerging topic detection or cascade spike prediction. In this paper, another new type of feature, namely competition, is proposed for task 2.

We conducted empirical evaluations on two sets of online movie reviews from two real social media sites. While our proposed ETP framework can be applied to most types of products, we take the task of predicting emerging topics for movies as the example to demonstrate and evaluate our proposed approach. The idea of taking the movie as the primary example comes from the studies from the electronic commerce community, which indicates that popularity of customer reviews has a greater impact on sales of experience products (like movies) than on those of search products (like commodities). The experimental results show that ETP delivers effective and efficient performance in predicting the emerging product topics from social media data.

The remaining of this paper is organized as follows: Section 2 introduces the background of emerging topic prediction. Section 3 gives the details of our ETP predictor. Experimental evaluation is discussed in Section 4. The last Section presents our conclusion.

2 RELATED WORK

Two primary areas related to our work are emerging topic detection, and cascades prediction over social media.

In the category of emerging topic detection, the existing works focused on topic representation and emerging detection mechanism. Most research represents a topic as a collection of words or hashtags [2, 6, 17]. For example, Cataldi et al. [2] presented the work that used aging theory to mimic the life cycle of each term and define emerging topics as sets of terms using a co-occurrence based metric. Emerging topics are detected by constructing keyword-based topic graph which connects the emerging terms with their co-occurrent ones under user-specified time constraints. The other approach for topic representation is developed based on topic modeling algorithms [1, 5, 14]. AlSumait et al. [1] proposed Online Latent Dirichlet Allocation to incrementally builds an up-to-date model when a new document appears. Saha et al. [14] proposed a dynamic non-negative matrix factorization framework with a complex temporal regularization. Hayashi [5] proposed a new method based on real-time streaming non-negative matrix factorization to detect top topic in twitter, and filter unrelated advertising tweets. These works are constrained in the detection of emerging topics, rather than prediction product topics in the early stage or before they are materialized.

In the category of cascade prediction over social media, an information cascade such as a photo, a hashtag, is considered to occur as information propagates through friends. Most work focused on predicting the popularity of the cascades [3, 8, 11]. Ma et al. [11] transformed the hashtag popularity range prediction problem into five-class classification problem. Content and social context features are investigated. The experimental results showed that social context feature are more effective than content feature. Kong et al. transform the hashtag popularity prediction problem into regression [8]. Among seven types of features, social structure is the 3rd effective feature, which is inferior to temporal and prototype feature. Given a cascade, Cheng et al. [3] developed a framework in predicting whether the cascade will continue to grow and double the cascade size in the future. They reported that structural and temporal and features are key predictors of cascade size. In particular, initially, breadth, rather than depth in a cascade is a better indicator of larger cascades. Wang et al. [16] investigated the burst (global spike) time prediction, rather than popularity prediction. They proposed a scale-independent classification-based approach. Experiments showed that fluctuation features are most important while social relation features and user profile features, which are Pagerank and HITS score of social network, are helpful.

All the cascade prediction research assumes that the complete network information is available. In particular, the experiments showed the effectiveness and superiority of social structure features. These problems have been solved in the aspect of range popularity or spike prediction on information cascades, but remain widely open on the emerging topic prediction with implicit network, which is shortfall of network structural information.

3 PROPOSED METHOD

3.1 Problem Statements

Briefly, the targeted research problem in this paper is early prediction of emerging products topics with implicit network. We will start with the definitions of topic, popularity and emerging.

Definition 1 (Time Window) *Time window* is the minimum time unit to measure popularity and features. It should be given before model learning and prediction. For example, the time window for emerging prediction of movie topic is one day.

Definition 2 (Topic Popularity) The *topic popularity tp*(*t*, *c*) of product topic *t* in the *c*-th time window, is the number of posts (product reviews) relevant to product topic *t* in the *c*-th time window.

Business people do not only look at the topic popularity of products, but also the novelty (significance) of popularity trends. There exist various definitions of emerging topic in the current literature due to different research goals. To assess the novelty of a popularity trend, we refer to the emerging score defined in [15].

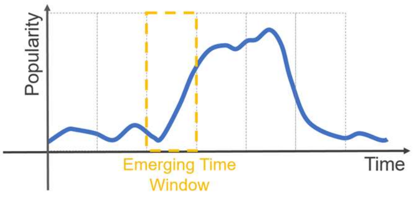

Figure 2: The Illustration of Emerging Product Topic

Definition 3 (Emerging Score) *Emerging Score ES*(*t*, *c*) of a topic *t* in the *c-th* time window is the measurement of the intensity of variation between two successive time windows,

$$ES(\boldsymbol{t},c)=\frac{tp(\boldsymbol{t},\mathrm{c})-EWMA(tp(t,1),tp(t,2),\ldots,tp(\boldsymbol{t},\mathrm{c}-1))}{1+EWMStd(tp(t,1),tp(t,2),\ldots,p(\boldsymbol{t},\mathrm{c}-1))}$$

where *EWMA* is exponentially weighted moving average, *EWMStd* is exponentially weighted moving standard deviation.

Definition 4 (Emerging Product Topic) Product topic *t* is emerging in the *c*-th time window, if *ES*(*t*, *c*) is larger than a specified threshold.

Definition 5 (Problem Statement of Task 1) Given a set of product review streams in social media site, the objective of task 1 is to predict which product topic *t* will emerge after current time window.

Task 1 is designed for long-term prediction in the pre-production stage. As illustrated in Figure 1, the ***observed time interval*** is the blue-shaded period where the solid curve is the current observed data while the dashed curve is the future data. Typically, the observed time interval of a product topic *t* for task 1 starts from the first time window of product topic *t*.

Definition 6 (Problem Statement of Task 2) Given a set of product review streams in social media site, the objective of task 2 is to predict which product topic *t* will emerge in the next ***predicted time interval***.

Task 2 is designed for short-term prediction after a product is launched. The *granularity* of the predicted time interval depends on the characteristics of a product. For example, in general, movies are released on Friday and the box office sales of movies are measured in terms of week. Therefore, the predicted time interval of task 2 for movies is set to the next week. Note that the observed time interval of task 2 is not the same as that of task 1 either. It is specified by the user. For example, film companies typically pay attention to the critical period which starts from two weeks before the movie is leased to four weeks after release. Therefore, as illustrated in Figure 1, on Monday of week 3 after release, a film company wishes to predict whether the movie will emerge in week 4. In this case, the observed time interval is the blue-shaded period while the predicted time interval is the 4th week after release.

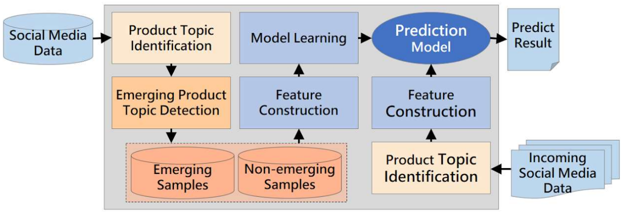

Figure 3: The proposed ETP Framework.

3.2 Framework of ETP (Emerging Topic Prediction)

The ETP predictor we proposed is shown in Figure 3. In offline phase, *Prediction Models* are constructed from the training data. First, *Product Topic Identification* module identifies the product topic of each post by named-entity extraction techniques. We assume that a dictionary of product names is available. For example, it is easy to get the movie titles and associated information from IMDB (http://www.imdb.com). Even with the dictionary of product names, it is not trivial to identify the product topic. Most movie names have alias. For example, the aliases appearing in reviews on the movie "The Fast and the Furious 6" includes of "Fast & Furious 6", "f&f6", "FAF6", "ff6", and so on. Besides, some reviews comment on multiple movies for comparison. The basic approach is to determine the product topic based on the frequencies of product names appearing in the post using the dictionary along with alias list.

Having identified the product topic of each training post, the *Emerging Product Topic Detection* module collects the stream of posts for each topic *t*, and for each time window *c*, assesses the emerging score *ES*(*t*, *c*) and determine whether product topic *t* is emerging in the *c*-th time window.

Model Learning module consists of a set of binary classification algorithms and constructs different prediction models for different length of observed time interval. For a long-term prediction model *LMl* of task 1, each stream of product *t* in the observed time interval $[tw_1^t, tw_l^t]$ is regarded as a training sample where tw_i^t is the *i*-th time window of product topic *t*. Those emerge after time window *l* are positive while others are negative samples. For a short-term prediction model *SMl* of task 2, each product stream in the designated observed time interval of length *l* is a training sample. Those emerge in the predicted time interval are positive while others are negative samples. As shown in Figure 1, the solid curves in the blue-shaded interval are the samples of observed length *l*. For each sample, *Feature Construction* module extracts four and five types of features for task 1 and task 2 respectively in terms of time window, which are described in the following sections.

Given incoming stream of product reviews, after *Product Topic Identification* and *Feature Construction*, for products in the pre-production stage with observed length *l*, long-term prediction model LM_l is employed to perform task 1 while for products in the opening weeks after launch, short-term prediction model SM_l is employed to perform task 2.

3.3 Author Diversity Feature

Social structure feature is widely utilized in most existing research on spike prediction of information cascades over social media [3, 8, 11]. Structure information of users and patterns of spreading paths to measure the broadness and depth of diffusion process are helpful for prediction of future trend. The rational behind social structure lies in that high viral cascades tend to spread across communities. For example, in [3], the density of the first-k re-share cascade is utilized to measure the tendency whether the diffusion is spreading across communities.

To predict emerging product topic in social media without social structure information, we proposed the author diversity features to assess the spreading tendency across different categories of users. The idea of the author diversity comes from the market segmentation in business. Market segmentation is the process of grouping consumers in the market into categories (known as segmentation) based on some types of shared characteristics. We categorize authors by preference, influence, engagement, and adoption respectively, and proposed four types of author diversities as follows. Note that the preference, influence, engagement, and adoption of an author are derived from the posting information of the author up to the current time window and are updated dynamically over time.

Definition 7 (Author's Preference) Given the number of posts $np_i(x)$ of genre i authored by *x*, the preference $prf_i(x)$ of author *x* for genre i is set to $np_i(x)$ in numerical mode, while in binary mode, $prf_i(x)$ is set to one if $np_i(x)$ is maximum among all genres and zero otherwise.

- preference of author x for genre i (binary mode)

$$prf_i(x) = \begin{cases} 1 & if\ i = arg \max_j(np_j(x)) \\ 0 & otherwise \end{cases}$$

- preference of author *x* for genre *i* (numerical mode)

$$prf_i(x) = np_i(x)$$

Definition 8 (Preference Diversity) The authors' preferences $q_i(c)$ for genre i in time window c is the summation of the preference $prf_i(x)$ for all authors x who have published posts during the time interval [1, *c*]. The preference diversity *PD*(*c*) in time window *c* is measured by entropy to quantify the impurity of authors in terms of preference of genres. *PD*(*c*) is defined as

$$PD(c) = -\sum_{\mathrm{i}=1}^{n} p_i\,(c)\log_2 p_i(c),$$

where

$$p_i(c) = {q_i(c)}\Big/{\sum_{i=1}^{n} q_i(c)},$$

$$q_i(c) = \sum_{t=1}^{\mathrm{c}} \sum_{x=1}^{m_t} prf_{\mathrm{i}}(\mathrm{x}),$$

c is the index of current time window, m_t is total number of authors in time window *t*, and *n* is total number of genres.

Definition 9 (Author's Influence) Given total number of responses $nr_i(x)$ with respect to all posts of genre *i* authored by *x*, the influence $inf_i(x)$ of author x on genre i is set to $nr_i(x)$ in numerical mode, while in binary mode, $inf_i(x)$ is set to one if $nr_i(x)$ is maximum among all genres and zero otherwise.

- influence of author *x* on genre *i* (binary mode)

$$inf_i(x) = \begin{cases} 1\ if\ i = arg \max_j(nr_j(x)) \\ 0\ \ else \end{cases}$$

- influence of author *x* on genre *i* (numerical mode)

$$inf_i(x) = nr_i(x)$$

Definition 10 (Influence Diversity) The cumulative authors' influence $q_i(c)$ on genre *i* up to the current time window *c* is the summation of the influence $inf_i(x)$ for all authors x who have published posts during the time interval [1, *c*]. The influence diversity *ID*(*c*) in time window *c* is measured by entropy in the following to quantify the impurity of authors in terms of influence of genres. *ID*(*c*) is defined as

$$ID(\mathrm{c}) = -\sum_{\mathrm{i}=1}^{n} p_i\,(c)\log_2 p_i,$$

where

$$q_i(c) = \sum_{t=1}^{\mathrm{c}} \sum_{x=1}^{m_t} inf_i(x),$$

$$p_i(c) = {q_i(c)}\Big/{\sum_{i=1}^{n} q_i(c)},$$

and definitions of *c*, m_t and *n*, are the same as those in Definition 5.

Definition 11 (Author's Engagement) The engagement of an author is measured by the gap (time interval) between two successive posts. In this paper, all the gaps are discretized into *el* levels. Given the number of posts $nr_i(x)$ with engagement level *i* authored by *x*, the engagement $eng_i(x)$ of author *x* for level *i* is set to $nr_i(x)$ in numerical mode, while in binary mode, $eng_i(x)$ is set to one if $nr_i(x)$ is maximum among all levels and zero otherwise.

- engagement of author *x* at level *i* (binary mode)

$$eng_i(x) = \begin{cases} 1\ if\ i = arg \max_j(nr_j(x)) \\ 0\ \ else \end{cases}$$

- engagement of author *x* at level *i* (numerical mode)

$$eng_i(x) = nr_i(x)$$

Definition 12 (Engagement Diversity) The cumulative authors' engagement $q_i(c)$ on level *i* up to the current time window *c* is the summation of the engagement $eng_i(x)$ for all authors *x* who have published posts during the time interval [1, *c*]. The engagement diversity *ED*(*c*) in time window *c* is measured by entropy in the following to quantify the impurity of authors in terms of engagement levels. *ED*(*c*) is defined as

$$ED(\mathrm{c}) = -\sum_{\mathrm{i}=1}^{el} p_i\,(c)\log_2 p_i(c)$$

where

$$q_i(c) = \sum_{t=1}^{\mathrm{c}} \sum_{x=1}^{m_t} eng_{\mathrm{i}}(x),$$

$$p_i(c) = {q_i(c)}\Big/{\sum_{i=1}^{n} q_i(c)},$$

and definitions of parameters are of the same as those in Definition 5 except that *el* is number of engagement levels.

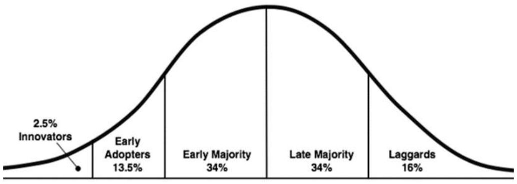

Figure 4: Adopters in Diffusion of Innovation[3]

According to Diffusion of Innovation Theory [13], originated in communication studies to explain the diffusion process of an idea or a product over time through a social system, some individuals are more apt to adopt the innovation (new product, or new idea) than others. There are five categories of adopters, innovators, early adopters, early majority, late majority and laggards (Fig. 4). Based on the posting order of a post in the corresponding topic stream *t*, it is easy to determine the category (adoption level) of the author in *t*.

Definition 13 (Author's Adoption) Given total number of posts $npa_i(x)$ for all posts of adoption level *i* authored by *x*, the adoption $adp_i(x)$ of author *x* on genre *i* is set to $npa_i(x)$ in numerical mode, while in binary mode, $adp_i(x)$ is set to one if $npa_i(x)$ is maximum among all genres and zero otherwise.

- adoption of author *x* at level *i* (binary mode)

$$adp_i(x) = \begin{cases} 1\ if\ i = arg\max_j(npa_j(x)) \\ 0\ \ else \end{cases}$$

- influence of author x at category i (numerical mode)

$$adp_i(x) = npa_i(x)$$

Definition 14 (Adoption Diversity) The cumulative authors' adoption $q_i(c)$ on level *i* up to the current time window *c* is the summation of the engagement $adp_i(x)$ for all authors *x* who have published posts during the time interval [1, *c*]. The adoption diversity *AD*(*c*) in time window *c* is measured by entropy in the following to quantify the impurity of authors in terms of adoption levels. *AD*(*c*) is defined as

$$AD(\mathrm{c}) = -\sum_{i=1}^{al} p_i\,(c)\log_2 p_i(c)$$

where $$q_i(c) = \sum_{t=1}^{c}\sum_{x=1}^{m_t} adp_i(x)$$

$$p_i(c) = {q_i(c)}\Big/{\sum_{i=1}^{n} q_i(c)}$$

and the definitions of parameters are of the same as those in Definition 5 except that *al* is number of adoption levels.

3.4 Competition Feature

Competition is a major principle of market economies. Competition occurs naturally among products in the same market. Products, which perform the same function, compete against each other under the budget constraints. For example, Fig. 5 shows the interplay of review popularities among three concurrent movies, Interstellar, Ghostbusters and Deadpool. It is therefore essential to take the competition nature into consideration for task 2. Recent studies have considered the competition effect for modeling of diffusion process. However, to the best of our knowledge, no work has incorporated the competition features into the prediction model for emerging topic or spike cascade prediction.

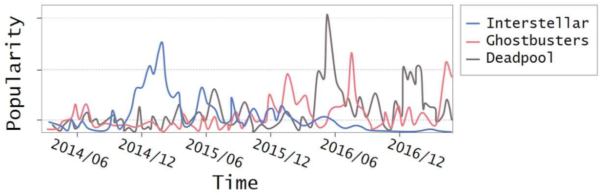

Figure 5: Popularity Interaction among Products

In this paper, we take the status of concurrent products into consideration. For a given target product, concurrent products are those whose launch periods up to now are overlapping with that of the target product (Sometimes, we may also shift the overlapping earlier to the pre-launch stage for promotion). We derive the following competition features.

- Number of concurrent products
- Number of concurrent products in terms of genre
- Number of emerging concurrent products up to now in terms of genre
- Number of emerging concurrent products over last week in terms of genre
- Correlation of review popularity between concurrent product and target product in terms of genre
- Number of time windows since the latest emerging event of concurrent product

Note that the correlation of popularity for a genre is the maximal Pearson Correlation among all concurrent products belonging to this genre.

3.5 Temporal, Content, and User Features

In addition to the proposed author diversity and competition features, we also derive three typical types of features as follows. All the following features are extracted in the unit of time window except that the product start time is a global feature.

Temporal features

- Total number of posts/responses till now
- Number of posts/responses in the last time window
- Average velocity of popularity
- Average acceleration of popularity
- Longest active time interval till now in terms of time window
- Longest inactive time interval till now in terms of time window
- Statistics of time interval between two posts/responses
- Number of emerging ones till now
- Number of time windows since the latest emerging event
- Start time of this product topic (Global)
- Latest time of this product topic
- Average popular time during a day
- Most popular day during a week
- Most popular month during a season
- Piecewise aggregate approximation of time between posts
- Statistics of 33.3% percentile of time between posts

Content features

- Average number of lines per post
- Average number of words per post
- Average sentiment score of a post

User features

- Total number of distinct authors till now
- Statistics of number of responses received by a author/responser

4 EXPERIMENTAL EVALUATION

4.1 Datasets and Experiment Setting

To evaluate the performance of our proposed approaches, we have performed experiments on real movie review data from Reddit (https://www.reddit.com) and PTT (https:/www.ptt.cc).

Reddit is the 4$^{\text{th}}$ most visit sites in U.S. According to Alexa Traffic Rank over the past 3 months, Reddit is ranked 7th in the world and 5th in U.S. As of Dec. 2017, Reddit has 234 million registered users, 50,000 active sub-reddits, 5 millions comments every day while the movies subreddit has more than 16 million subscribers. The Reddit dataset comes from the Datasets subreddit. We extract the posts along with comments in the Movie subreddit from January 2013 to December 2016. There are 867,468 posts, 15,442,009 comments, 1,205,248 users (including authors and responders). Posts containing *url* only are filtered out first. In the end, 711 movies are identified, after product topic identification with the movie dictionary crawled from IMDB.

PTT is the largest bulletin board system in Taiwan. According to Alexa Traffic Rank over the past 3 months, PTT is ranked 15th in Taiwan. PTT has 1.5 million registered users, over 20,000 boards, 20,000 articles and 0.5 million comments every day. The PTT dataset is collected from our developed crawler. There are 117,328 posts, 3,554,029 comments, 154,307 users and 486 movies are identified. The movie dictionary, which contains Chinese movie names and release dates in Taiwan, is crawled from TrueMovie (http://www.truemovie.com).

We use five-fold cross-validation to evaluate on accuracy, precision, and recall for emerging topic prediction. Experiments are performed to evaluate the prediction performance versus observed length, classification algorithm, and features.

4.2 Performance for Task 1

Figure 6 shows the accuracy as a function of observed length for comparison of various classification algorithms, Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM) and Gradient Boosting Decision Tree (GBDT), on Reddit. Random Forest performs best with accuracy over 90% except the first time window. The accuracy of all algorithms is better than 85% in spite of the observed length. It is no surprise that the performance improves approximately with increasing observed length except some time intervals.

To explore the prediction performance more in depth, Table 1 lists the earliness, accuracy, precision, and recall for various observed length on Reddit using Random Forest.

Definition 15 (Earliness for Task 1) The earliness k_l^t of a long-term prediction model LM_l with respect to a product topic t is the number of time windows from time window l to the first emerging time window.

Earliness depends on the observed length and the emerging time (Figure 1). Intuitively, the longer the observed length, the better the prediction accuracy, but the shorter the earliness. From Table 1, it can be seen that the proposed prediction model for task 1 achieves promising accuracy of at least 91% as early as 285 days in average before the first emerging event. Moreover, the accuracy could be improved by improving the recall of emerging topics.

As most research on prediction over social media reported that temporal and social structure-related types of features are key indicators [3, 8, 11], Figure 7 examines the effect of proposed author diversity features against other types of features. The proposed diversity features surprisingly perform not only better than other features but also near the full set of features. This implies that the proposed author diversity is a well-performed predictor for prediction with implicit network.

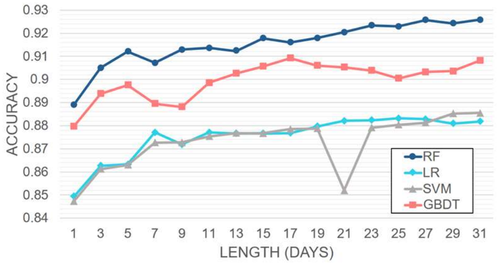

Figure 6: Accuracy versus Observed Length of Different Algorithms of Task 1 in Reddit.

Table 1. Earliness, Precision and Recall versus Observed Length using Random Forest of Task 1 in Reddit.

Length	Earliness (Days)	Accuracy	Emerging		Non-Emerging	
			Precision	Recall	Precision	Recall
1	293.532	0.888886	0.799	0.367	0.896	0.983
3	291.532	0.905022	0.868	0.412	0.909	0.988
5	289.532	0.912066	0.893	0.437	0.914	0.991
7	287.532	0.907235	0.850	0.427	0.913	0.987
9	285.532	0.912942	0.899	0.446	0.915	0.991
11	283.532	0.913653	0.912	0.420	0.915	0.993
13	281.532	0.912306	0.904	0.409	0.915	0.991
15	279.532	0.917842	0.914	0.431	0.920	0.992
17	277.532	0.915921	0.921	0.401	0.916	0.994
19	275.532	0.917961	0.897	0.421	0.920	0.993
21	273.532	0.920382	0.918	0.419	0.921	0.994
23	271.532	0.923339	0.926	0.434	0.924	0.994
25	269.532	0.923020	0.907	0.436	0.924	0.994
27	267.532	0.925717	0.913	0.439	0.927	0.994
29	265.532	0.924268	0.931	0.402	0.924	0.996
31	263.532	0.925810	0.902	0.433	0.928	0.993

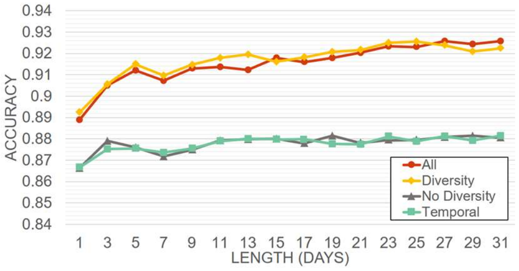

Figure 7: Accuracy versus Observed Length of Different Features using Random Forest of Task 1 in Reddit.

Figure 8, Table 2 and Figure 9 show the results of the same experiments on PTT. Again, Random Forest performs best, but the prediction accuracy is better than that on Reddit. By observing Table 2 where *ETP* on PTT achieves 95% accuracy as early as 104 days before the first emerging event, the reason may come from shorter earliness of PTT comparing to that of Reddit. Shorter earliness tends to predict better. Users of PTT from Taiwan are less apt to access movie news in the early stage of pre-production from Hollywood than those in Reddit. Moreover, Figure 9 also reveals that the proposed author diversity is a good predictor. But the superiority is not as promising as that on Reddit. This can be explained by inspecting the author diversities between Reddit and PTT. In average, authors of Reddit are more diverse than those of PTT.

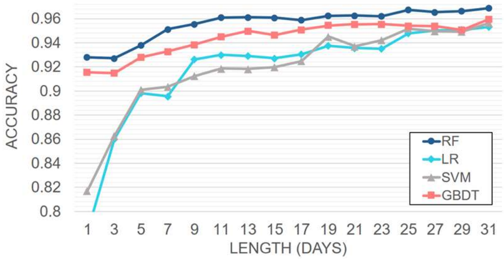

Figure 8: Accuracy versus Observed Length of Different Algorithms of Task 1 in PTT.

Table 2. Earliness, Precision and Recall versus Observed Length using Random Forest of Task 1 in PTT.

Length	Earliness (Days)	Accuracy	Emerging		Non-Emerging	
			Precision	Recall	Precision	Recall
1	110.257	0.927874	0.852	0.844	0.953	0.953
3	108.257	0.927226	0.840	0.839	0.955	0.952
5	106.257	0.938029	0.858	0.856	0.962	0.960
7	104.257	0.951381	0.895	0.873	0.968	0.972
9	102.257	0.955442	0.908	0.868	0.968	0.977
11	100.257	0.960944	0.927	0.869	0.970	0.983
13	98.257	0.961348	0.920	0.878	0.972	0.981
15	96.257	0.960863	0.914	0.878	0.972	0.980
17	94.257	0.958883	0.920	0.858	0.969	0.982
19	92.257	0.962407	0.933	0.858	0.970	0.985
21	90.257	0.962707	0.931	0.857	0.970	0.986
23	88.257	0.961898	0.930	0.851	0.969	0.986
25	86.257	0.967294	0.942	0.858	0.973	0.989
27	84.257	0.965555	0.934	0.853	0.972	0.987
29	82.257	0.966328	0.941	0.850	0.972	0.989
31	80.257	0.968735	0.949	0.857	0.973	0.990

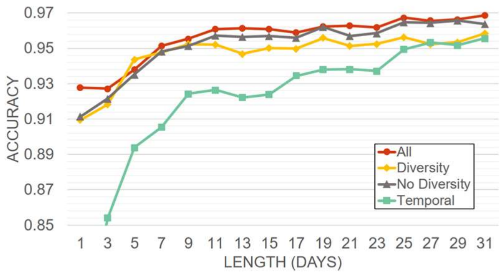

Figure 9: Accuracy versus Observed Length of Different Features using Random Forest of Task 1 in PTT.

4.3 Performance for Task 2

While Task 2 is designed for short-term emerging prediction on next predicted time interval, in our experiment, the granularity of predicted time interval is set as one week while the observed time interval is designated as the period starting from two weeks before a movie is released till current time window (Figure 1). Figure 10 shows the prediction accuracy as a function of observed length for comparison of various classification algorithms on Reddit. Note that x-axis denotes the observed length and leads to the predicted time interval implicitly. For example, the observed length 24 indicates that, on the 3rd day of week 2 after movie release, the user wishes to predict whether the movie will emerge in week 3. In Figure 10, Random Forest still performs best with accuracy over 96% except the prediction on opening week with accuracy around 92%. This may due to the promotional campaigns prior to the opening week.

Table 3 lists the precision, and recall versus difference earliness for the prediction on the first three opening weeks using Random Forest on Reddit.

Definition 16 (Earliness for Task 2) The earliness k_l^t of a short-term prediction model SM_l with respect to a product topic t is the number of time windows from the last time window of observed time interval to the first time window of predicted time interval.

For example, if the observed length is 24, the earliness is 4 days before next Friday (Recall that most movies released on Friday). From Table 3, it can be observed that the performance can be improved by improving the recall of emerging ones.

Figure 11 examines the effect of competition feature on Reddit using Random Forest. Once again, the performance of the proposed competition feature alone is no less inferior to the full set of features. This indicates that competition feature is a good predictor for Task 2 on Reddit.

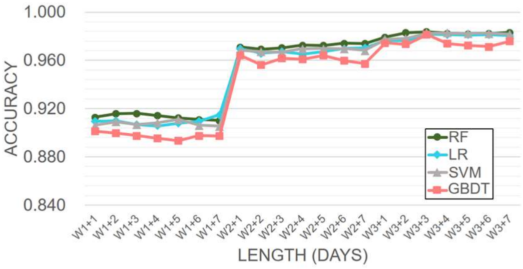

Figure 10: Accuracy versus Observed Length of Different Algorithms of Task 2 in Reddit.

Table 3. Earliness, Precision and Recall versus Earliness for the First Three Opening Weeks using Random Forest of Task 2 in Reddit.

k	Predict Week 1				Predict Week 2				Predict Week 3			
	Emerging		Non-emerg		Emerging		Non-emerg		Emerging		Non-emerg	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
7	.632	.224	.923	.986	.673	.242	.974	.997	.340	.120	.981	.999
6	.678	.242	.925	.987	.633	.224	.972	.997	.520	.200	.984	.999
5	.670	.209	.924	.989	.620	.240	.974	.996	.300	.147	.985	.999
4	.632	.209	.924	.987	.657	.248	.975	.997	.160	.060	.984	.999
3	.647	.206	.924	.985	.707	.256	.975	.997	.180	.087	.984	.998
2	.606	.197	.922	.985	.643	.276	.977	.997	.160	.067	.984	.999
1	.590	.190	.921	.985	.746	.296	.977	.996	.180	.080	.984	.999

Figure 12, Table 4 and Figure 13 show the results of the same experiments on PTT. Again, Random Forest performs best, but the prediction accuracy of the first open week is inferior to that on Reddit while the 2nd and 3rd opening week are superior. It can be observed that the proposed approach on PTT achieves more than 95% accuracy as early as one week before for emerging prediction on next week except the first opening week. Figure 13 examines the effect of competition features on PTT. The competition feature alone on PTT is slightly worse than the full set feature. Though the performance of competition feature is not as good as that on Reddit, it is helpful to raise the accuracy on PTT.

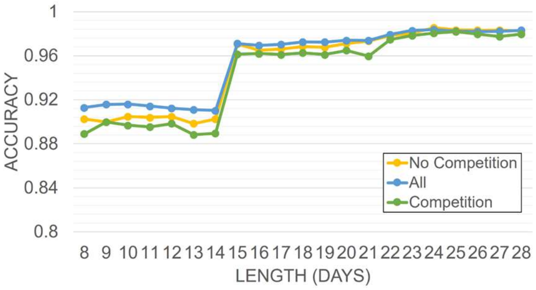

Figure 11: Accuracy versus Observed Length of Different Features using Random Forest of Task 2 in Reddit.

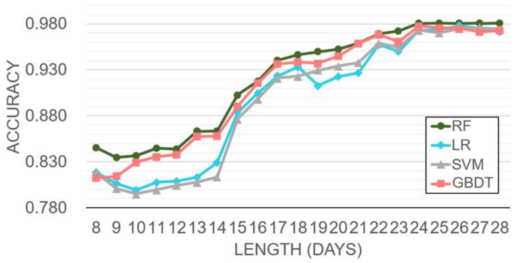

Figure 12: Accuracy versus Observed Length of Different Algorithms of Task 2 in PTT.

Table 4. Earliness, Precision and Recall versus Earliness for the First Three Opening Weeks using Random Forest of Task 2 in PTT.

k	**Predict Week 1**				**Predict Week 2**				**Predict Week 3**			
	Emerging		Non-emerg		Emerging		Non-emerg		Emerging		Non-emerg	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
7	.788	.454	.876	.967	.905	.536	.962	.995	.733	.453	.983	.998
6	.806	.436	.873	.971	.842	.484	.958	.993	.743	.440	.983	.998
5	.711	.387	.862	.959	.858	.461	.955	.992	.827	.433	.983	.998
4	.721	.361	.859	.965	.850	.437	.952	.991	.767	.427	.983	.998
3	.687	.380	.858	.954	.812	.370	.946	.992	.720	.420	.983	.998
2	.674	.350	.856	.955	.800	.331	.925	.989	.828	.450	.976	.995
1	.697	.302	.859	.970	.801	.385	.912	.984	.775	.564	.977	.991

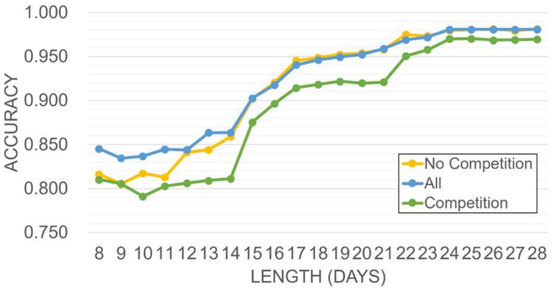

Figure 13: Accuracy versus Observed Length of Different Features using Random Forest of Task 2 in PTT.

4 CONCLUSIONS

In this paper, we have addressed the problem of emerging product topic prediction over social media with implicit networks. Two prediction tasks are investigated. One is the long-term prediction in the pre-production stage to predict which products topics will emerge in the future while the other is the short-term prediction after product launch to predict which products topics will emerge in the next predicted time interval. A novel framework named *ETP* is developed to deal with the two targeted tasks. Moreover, two novel features named author diversity and competition, are proposed to deal with the diffusion process without social structure information and the short-term prediction task, respectively. Experiments performed on two real movie reviews datasets from Reddit and PTT show that for long-term prediction on Reddit, ETP achieves promising accuracy of at least 91% as early as 285 days before the first emerging event while 95% accuracy as early as 104 days before the first emerging on PTT. In particular, the proposed author diversity features surprisingly perform not only better than other features but also near the performance of full set of features. For short-term prediction, ETP performs well with accuracy over 96% for prediction on the 2nd and 3rd opening week while 92% on the first opening week. Moreover, the performance of the proposed competition feature alone is no less inferior to the full set of features. In summary, the proposed ETP framework along with the novel author diversity and completion features are shown to constitute a novel predictor for emerging product topic prediction over social media with implicit networks.

ACKNOWLEDGEMENT

This research was partially supported by Ministry of Science and Technology, Taiwan, under grant number 106-3114-E-009-008.

REFERENCES

[1] L. Alsumait, D. Barbará, and C. Domeniconi, On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking, 8th IEEE International Conference on Data Mining, 2008.

[2] M. Cataldi, L. D. Caro, and C. Schifanella, Emerging topic detection on Twitter based on temporal and social terms evaluation, Proceedings of the 10th International Workshop on Multimedia Data Mining, 2011.

[3] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, Can cascades be predicted?, Proceedings of the 23rd International Conference on World Wide Web, 2014.

[4] W. Duan, B. Gu, and A. B. Whinston, The dynamics of online word-of-mouth and product sales: An empirical investigation of the movie industry, Journal of Retailing, Vol. 84, No. 2, 2008.

[5] K. Hayashi, T. Maehara, M. Toyoda, and K.-I. Kawarabayashi, Real-time Top-R topic detection on Twitter with topic Hijack filtering," Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[6] S. Huang, Y. Yang, H. Li, and G. Sun, Topic detection from microblog based on text clustering and topic model analysis, Asia-Pacific Services Computing Conference, 2014.

[7] J. Hurtado, S. Huang, and X. Zhu, Topic discovery and future trend prediction using association analysis and ensemble forecasting, IEEE International Conference on Information Reuse and Integration, 2015.

[8] S. Kong, Q. Mei, L. Feng, F. Ye, and Z. Zhao, Predicting bursts and popularity of hashtags in real-time, Proceedings of the 37th international ACM SIGIR Conference on Research & Development in Information Retrieval, 2014.

[9] J. Li, G. Dong, and K. Ramamohanarao, Making use of the most expressive jumping emerging patterns for classification, Knowledge Discovery and Data Mining. Current Issues and New Applications Lecture Notes in Computer Science, pp. 220–232, 2001.

[10] Y. Liu, Word of mouth for movies: Its dynamics and impact on box office revenue, Journal of Marketing, Vol. 70, No. 3, 2006.

[11] Z. Ma, A. Sun, and G. Cong, On predicting the popularity of newly emerging hashtags in Twitter, Journal of the American Society for Information Science and Technology, Vol. 64, No. 7, pp. 1399–1410, 2013.

[12] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowledge-Based Systems, Vol. 69, pp. 24–33, 2014.

[13] E. M. Rogers, Diffusion of innovations. London: Simon & Schuster, 2003.

[14] A. Saha and V. Sindhwani, Learning evolving and emerging topics in social media, Proceedings of the fifth ACM international conference on Web Search and Data Mining, 2012.

[15] E. Schubert, M. Weiler, and H.-P. Kriegel, SigniTrend, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.

[16] S. Wang, Z. Yan, X. Hu, P. S. Yu, Z. Li, and B. Wang, CPB: a classification-based approach for burst time prediction in cascades, Knowledge and Information Systems, Vol. 49, Issue. 1, 2016.

[17] S.-H. Yang, A. Kolcz, A. Schlaikjer, and P. Gupta, Large-scale high-precision topic modeling on twitter, Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014.

[18] R. Zafarani, M. A. Abbasi, and H. Liu, Social media mining: an introduction. New York: Cambridge University Press, 2014.