Extended Version: Effective Candidate Selection and Interpretable Interest Extraction for Follower Prediction on Social Media

Seiji Maekawa
maekawa.seiji@ist.osaka-u.ac.jp
Osaka University
Osaka, Japan

Santi Saeyor
santi@hottolink.co.jp
Hottolink Inc.
Tokyo, Japan

Takeshi Sakaki
t.sakaki@hottolink.co.jp
Hottolink Inc.
Tokyo, Japan

Makoto Onizuka
onizuka@ist.osaka-u.ac.jp
Osaka University
Osaka, Japan

ABSTRACT

We address the problem of predicting new followers for a company account given that the number of social network API calls is limited, in order to enhance the marketing communication effectiveness on social media. The contributions of this paper are three-fold: 1) filtering methods that select promising candidate accounts with high precision while effectively reducing the number of API calls, 2) a new method for extracting interpretable feature vectors (interest vectors) from each account by utilizing standardized categories for marketing communication, and 3) a follower prediction model by utilizing the above candidate selection methods and interpretable interest vectors. Experiments on Twitter data confirm that our follower prediction model performs well with a small number of API calls and clarify which dimension of interest vectors (interest category) contributes to prediction performance.

CCS CONCEPTS

• **Information systems** → **Social networks**.

KEYWORDS

social media, API restrictions, link prediction, interpretability

ACM Reference Format:
Seiji Maekawa, Santi Saeyor, Takeshi Sakaki, and Makoto Onizuka. 2021. Extended Version: Effective Candidate Selection and Interpretable Interest Extraction for Follower Prediction on Social Media. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI'21).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Social media, e.g., Instagram and Twitter, are important touchpoints for companies in terms of marketing and making public relations with users. To enhance brand awareness and widely disseminate their posts on social media, companies need to increase the number of followers; their posts will reach a greater number of potential customers and receive increased positive reactions, e.g., likes and shares. It is known that these reactions enhance sale volume [8].

There are two major approaches for gaining more followers on social media, which are direct and indirect approaches. As for the direct approach, advertising tools [25] provided by Instagram and Twitter permit advertisers to specify target user segments based on user's location, gender, language, interest, keyword, custom audience, etc. As for the indirect approach, there are strategies that utilize the nature of social media, such as viral marketing [12] that utilizes information diffusion among user accounts and influencer marketing [27] that utilizes the existence of influencers commonly followed by prospective followers. Both of direct and indirect approaches utilize posts, profile, and follow-relationships of user accounts, which are the major types of information associated with user accounts on social media.

Since there are several choices for specifying user segments in the direct approach and selecting strategies in the indirect approach, marketers have two tasks in order to make the appropriate choices; 1) they need to predict user accounts that are likely to follow a given company account (we call this set of those user accounts ***predicted account set***) and 2) they need to clarify which characteristics of user accounts contribute to prediction results since they need to explain to their clients how the account set was predicted.

Requirements To build practical marketing communication services, we focus on two requirements that have not been considered largely in research communities. The first requirement is that we need to effectively extract predicted account sets with a small number of social network application programming interfaces (APIs) calls to collect users' posts, profile, and follow-relationships. This is because the number of permitted API calls per minute is limited, e.g., Twitter allows us to invoke a single API call per minute to collect the follower IDs of a specified account (we describe the details in Section 3.1). The second requirement is the interpretability of feature vectors of user accounts. Marketers need the interpretability in order to correctly analyze the prediction results. The interpretability of feature vectors makes it possible to investigate which characteristics of user accounts are important to predict new followers.

WI'21, Dec. 2021, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Actually, interpretability of prediction results is a hot research field in various domains [10]. However, we observe that the interpretability of feature vectors is not a popular topic in research communities. In particular, no interpretable feature vector has been proposed for marketing communication.

Existing methods As for the first requirement, to the best of our knowledge, no study has yet to target the identification of predicted account sets under the limited number of API calls. Although link prediction techniques [28, 33] are related to the identification, they cannot be directly applied since they need to access all data on graphs, which cannot be collected due to the restriction of API calls. As for the second requirement, conversion from posts and profile to feature vectors is widely studied (e.g., word embedding [5, 19] and sentence embedding [3, 11]). However, the interpretability of those feature vectors is not high [13].

Contributions We address the link prediction problem for effectively extracting predicted account sets by utilizing interpretable feature vectors under the limited number of API calls. We call this problem ***follower prediction***. The contributions of this paper are three-hold. First, we propose candidate account selection methods utilizing user attributes and social relationships in order to reduce the number of API calls. The basic idea is filtering user accounts by appropriately generating filtering conditions on their posts (user attributes) or follow-relationships (social relationships), which are useful for follower prediction. We clarify the applicability of those methods by using real-world data. Second, we propose a method to extract interpretable feature vectors by calculating correlations between advertising categories and user accounts from their text information. We use standardized categories [7] as the feature vector dimensions of user accounts so that marketers can interpret the feature vectors as the users' interests that we call ***interest vectors***. Finally, we construct a follower prediction model by utilizing the above candidate account selection methods and the interpretable interest vectors. We validate the effectiveness of our proposal by comparing it with various existing methods using real data from Twitter. Furthermore, we show use cases that the interests related to a given company actually contribute to follower prediction.

Organization The rest of this paper is organized as follows. Section 2 describes the details of related work. We describe an overview of our proposal in Section 3. Section 4 proposes three candidate account selection methods and investigates their effectiveness. Section 5 proposes our interest extraction method. We construct a follower prediction model in Section 6. Section 7 gives the purpose and results of evaluations and a case study. Finally, we conclude this paper and discuss the limitations of our proposal in Section 8.

2 RELATED WORK

First, we discuss the existing studies on the two requirements mentioned in Section 1: 1) candidate account selection with limited data access in Section 2.1 and 2) interpretable feature vectors of accounts on social media in Section 2.2. Then, we explain existing link prediction techniques related to follower prediction in Section 2.3.

2.1 Methods Considering Limited Data Access

Many studies [20, 23] have proposed methods to estimate the topological structure of social networks under the restriction of APIs. Smith and Thai [23] efficiently calculate the strength of the connections between the nodes by using the random walk technique. Nakajima and Shudo [20] estimate the size of the entire graph from restricted data access by using biased random walk approach considering private nodes. To the best of our knowledge, no follower prediction model has considered strict limits on data availability.

Local community detection methods [4, 15], which identify the community to which a query node belongs, have been proposed. These methods access only partial nodes by utilizing metrics such as k-core and k-truss [4]. However, it is impractical to collect follow-relationships between all nodes close to company accounts via API calls because they are generally high-degree nodes. They are usually adjacent to over tens of thousands of nodes within 1-hop and adjacent to millions of nodes within 2-hops. Therefore, we need a truly effective candidate account selection method.

2.2 Interpretable Feature Vector

Feature extraction methods for objects with text information have been widely studied, such as word embedding and sentence embedding [5, 11]. However, the feature vectors of word embedding and sentence embedding are insufficiently interpretable.

With the spread of social media in recent years, interest extraction has been widely studied for user modeling for web mining [21, 30]. There are two major approaches to interest extraction; 1) supervised learning methods, e.g., text classification and document labeling [14, 22], and 2) unsupervised learning methods, e.g., topic modeling, and document clustering [6, 32]. First, supervised learning methods have the disadvantage that continuously creating training data requires a significantly high cost due to the real-time nature of social media. Second, unsupervised methods do not use training data. However, most unsupervised learning methods extract insufficient interpretable features. Some other methods extract easily interpretable features of input documents because they use preset categories as labels for documents. For example, the methods proposed in [9, 24] utilize Wikipedia as a knowledge base or an ontology in order to label input documents.

In particular, the method proposed in [24] is suitable for building highly expressive features because it supports more than 100 labels while other methods support only a few dozen labels. However, this existing method has several problems in constructing an ontology-based network from Wikipedia and the limitations of using Wikipedia as a data source for marketing communication.

2.3 Link Prediction

Link prediction can be considered as follower prediction by treating accounts as nodes and follow-relationships as edges [29]. Several metrics that can be used to represent the topological similarity between nodes have been widely studied, such as Jaccard index and Adamic/Adar [1]. Recently, deep learning-based methods have been proposed [28, 33]. SEAL [33] achieves high accuracy in link prediction by learning the topological information conveyed by the edges and attributes associated with the nodes.

Regarding use on social media, existing methods have a drawback in that they need to access all data on graphs and do not consider restrictions that the number of permitted API calls per minute is

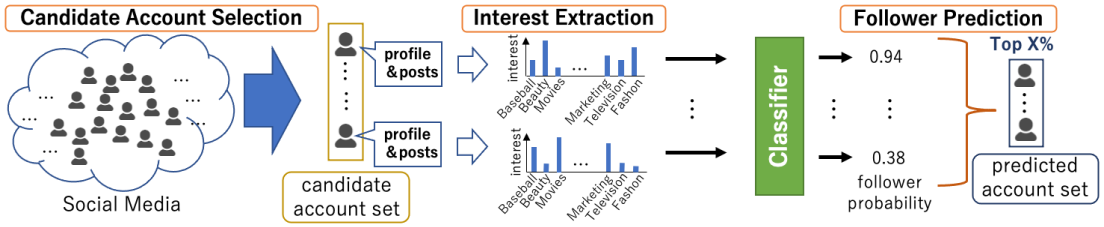

Figure 1: Overview of Our Approach.

limited. Therefore, we need a new approach to predict accounts that are likely to follow a company account under this restriction.

3 OVERVIEW OF PROPOSAL

We first explain challenges in solving follower prediction problem. Second, we describe a problem definition of predicting user accounts that are likely to follow a given company. Finally, we describe an overview of our proposal.

3.1 Challenges

There are two major challenges in solving follower prediction problem as follows. The first challenge is that we need to design our follower prediction approach that should be effective under the limited number of API calls. For instance, Twitter allows us to invoke a single API call per minute at most to collect the IDs of the followers of a specified account. A single call is also restricted to return 5000 followers at most. Therefore, at least one million minutes (approximately two years) are required to collect all the followers of one million accounts. It is impractical to collect all accounts in a whole social network because the monthly active users on Twitter number billions of accounts globally, with 40 million accounts in Japan[1]. The second challenge is that we need to design the feature vectors of accounts such that the vectors are interpretable in order to correctly analyze the prediction results.

3.2 Problem Statement

We give the problem definition below:

Definition. Given a company account, we predict new follower accounts (predicted accounts) that are likely to follow the company account in the next time step (e.g., next month), satisfying two conditions; 1) we invoke a small number of API calls and 2) feature vectors used for prediction are interpretable.

A network on social media is expressed as a graph $G = (V, E)$, where $V = \{v_1, \ldots, v_l\}$ represent accounts, $E \subseteq \{(v_i, v_j) | v_i, v_j \in V\}$ represent account v_i follows account v_j, and l is the total number of accounts on social media. In addition, we extract feature vectors from the profile and posts of each account and attach them to each node as attributes.

3.3 Approach

To tackle the first challenge, we take a three-step approach that extracts predicted account sets under the limited number of API calls. We show its overview in Figure 1. In the first step, we select candidate accounts $V_{cand} \subset V$ heuristically with a small number of API calls, where V_{cand} is a set of accounts that are likely to follow a given company account $v_c \in V$. For this candidate selection, we generate filters based on the attributes or topology (posts or follow-relationships) of user accounts, which are useful for follower prediction (details are given in Section 4). In the second step, we extract the interests of those accounts from their profile and posts as their feature vectors. In the third step, we calculate the probability that each $v_i \in V_{cand}$ newly follows v_c in the next time step. By extracting top $X\%$ accounts based on the probability, we can obtain the predicted account set of the company account v_c.

To tackle the second challenge, we propose a method to extract interests (expressed as interpretable feature vectors) of accounts on social media (this is detailed in Section 5). Because our focus is on marketing communication, we employ standard advertising categories provided by the International Advertising Bureau Categorization (IAB Categorization) [7] and adopt them as the dimensions of feature vectors representing the interests of accounts.

Table 1: Summary of Datasets. Com. indicates company.

Company	Industry domain	# of followers	Follow on Aug.	Follow on July	Follow on June
com. A	Fashion	3.68e4	1703	1766	1616
com. B	Food	5.11e4	9388	3135	3631
com. C	Cosmetic	3.11e5	23418	23634	16664
com. D	Restaurant	7.66e5	46751	66349	27684

4 CANDIDATE ACCOUNT SELECTION

In this section, we first propose three candidate account selection methods which can be executed with a small number of social network API calls. Second, we compare those methods to evaluate their quality, using real data of four company accounts collected on Twitter. Table 1 shows the summary of the data.[2] The four company accounts were randomly chosen from different industries and have different numbers of followers.

4.1 Selection Approaches

We propose three selection approaches as candidate account selection methods, which are simple and effective for practical marketing communication. Figure 2 shows an overview of the three candidate account selection methods.

The first approach is keyword-based selection that selects candidate accounts that posted tweets mentioning keywords approved by the given company. Since we can only collect the current tweets from the API calls, we continuously collect tweets and select those

[1]Note that Twitter does not permit us to register multiple applications for a single use case, or substantially similar or overlapping use cases [26].

[2]In the interest of privacy, we keep the companies anonymous to protect their information.

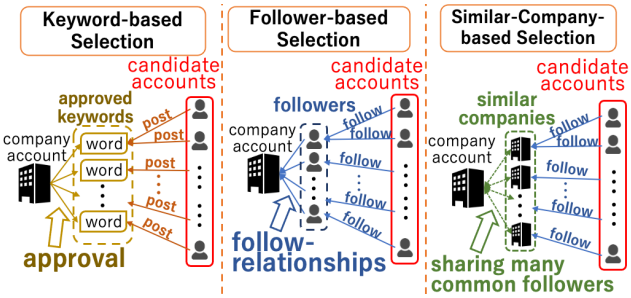

Figure 2: Overview of Candidate Account Selection Methods.

Table 2: Precision Ratios (Per Ten Thousand Accounts) of New Followers to Candidate Accounts. In parentheses, we show the number of candidates.

Company	Keyword-based selection	Follower-based selection	Similar company -based selection
Company A	0.8 (7.4e6)	2.9 (0.7e6)	2.3 (2.4e6)
Company B	7.1 (5.0e6)	12.9 (0.9e6)	9.0 (5.5e6)
Company C	13.7 (1.5e6)	15.0 (4.3e6)	23.9 (0.8e6)
Company D	24.2 (5.1e6)	18.1 (13.0e6)	46.6 (5.8e6)
Average rank	2.75	1.75	1.5

candidate accounts during a specified period, e.g., the last month. This approach is effective when the keywords are embedded into discriminative features used for follower prediction. The second approach is follower-based selection that selects candidate accounts that follow the recent followers of the given company. This approach leverages 1) second-order (2-hop) proximity between the candidates and the given company and 2) the recency of follow-relationships between them. The reason we use recent followers is that we assume the recent followers are more informative in predicting new followers than other existing followers[3]. The third approach is similar-company-based selection that selects user accounts that follow company accounts sharing many followers with the given company. The idea of this approach is to employ collaborative filtering that utilizes follow-relationships of similar company accounts. This approach is effective when similar company's follow-relationships play an important role for follower prediction.

These candidate account selection methods can be realized with fewer API calls than collecting all posts and follow-relationships. This is because 1) posted tweets mentioning specified keywords are significantly fewer than all tweets, 2) follow-relationships of the recent followers are also significantly fewer than all follow-relationships in Twitter, and 3) follow-relationships of the similar companies are also significantly fewer than all follow-relationships.

4.2 Comparison of Selection Approaches

We validate the effectiveness of the three candidate account selection methods by using real data. We predict new followers to be included in candidate account sets. In this experiment, we treat the followers acquired in August 2020 as new followers and select candidate account sets from data acquired before August 2020. Regarding the keyword-based selection approach, we use the keywords that were approved by the companies and actually used in advertising campaigns. We identify the accounts that posted tweets mentioning the keywords related to the given company in July 2020. For the follower-based selection approach, we identify the accounts that follow the given company account's followers acquired in July 2020. Finally, for the similar-company-based approach, we select company accounts that have high Jaccard index scores with the given company in terms of their follower sets.

Table 2 shows the precision ratios which indicate the ratios of new followers to the candidate accounts selected by the selection approaches. In the last row shows the average rank of each selection approach for all companies. The results indicate that both the follower-based selection approach and similar-company-based selection approach achieve higher precision ratios than the keyword-based selection approach. Consequently, we conclude that these two approaches can effectively select candidate account sets with high precision ratio under the limited number of API calls.

5 INTEREST EXTRACTION

In this section, we propose a method for extracting the interests of user accounts by extending the existing method using spreading activation on an ontology-based network. We utilize Wikipedia for constructing ontologies since Wikipedia is continuously updated and contains new words and concepts, which is suitable for marketing communication. First, we define account interests to ensure high interpretability for marketers. Second, we explain the problems of the existing method and extensions to solve them. Third, we present the specific steps of our proposal.

5.1 Interest Definition

In proposing a method for extracting account interests, we first define what is meant by interest. As mentioned in Section 2.2, account features must be interpretable because marketers need to know the interest of each account to clarify which characteristics of user accounts contribute to prediction results. To obtain interpretable account feature vectors, we utilize preset interest categories, IAB Categorization, as the dimensions of feature vector. The IAB Categorization is one of the most general and standard categorizations used to classify advertising contents and they are easy for marketers to interpret. This categorization is also used in Twitter's SNS targeted advertisement method[4].

5.2 Comparison of Existing and Proposed Methods

We adopt Syed et al.'s method, which applies spreading activation, an unsupervised learning method, to document labeling and uses an ontology-based network constructed from Wikipedia [24]. Since this method clarifies which Wikipedia categories are strongly correlated with input documents by spreading activation, we can obtain the interpretable results of document labeling.

Spreading activation is a technique of searching for nodes correlated with input documents on various networks such as semantic networks. It works as follows: give a weight as activation values to the initial group of article nodes of the network corresponding to

[3]We validate this assumption by conducting experiments and show the results in Appendix A of the extended version of this paper [16] due to the space limitation.

[4]https://developer.twitter.com/en/docs/twitter-ads-api/campaign-management/api-reference/iab-categories

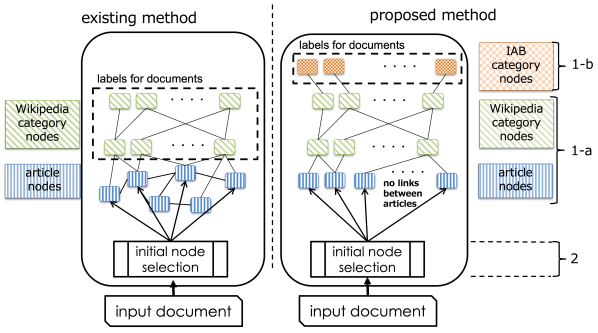

Figure 3: Existing Method vs. Proposed Method.

the results of search query; propagate activation values on the network iteratively for a certain number of times or until the weights of all nodes become stable; choose category nodes with activation value greater than 0 as the search result. This method consists of the following three steps.

1. Ontology-Based Network Construction: We construct a network using the Wikipedia link structure as an ontology. In the network, we use Wikipedia article and Wikipedia category as nodes, and create edges between node pairs with inter-article links, inter-category links, and inter-article-category links.

2. Initial Node Selection: We search for Wikipedia article nodes in the ontology-based network by using an input document as a query and set top k article nodes of the search results as the initial nodes to be activated.

3. Spreading Activation: We perform spreading activation from the initial nodes. Among the activated nodes, category nodes with activation values greater than 0 are considered as the Wikipedia categories to match the input document.

There are two problems with the existing method. First, the existing method cannot be directly applied to interest extraction of accounts for marketing communication on social media, since the method uses all Wikipedia categories and most of those categories are not related to marketing communication. Second, Wikipedia category nodes, which are document labels, are not sufficiently activated and documents are not labeled appropriately. This is because Wikipedia article nodes which are more reachable from many Wikipedia article nodes tend to be activated more strongly than Wikipedia category nodes.

To solve these problems, we extend the Ontology-Based Network Construction step. To deal with the first problem, we add IAB categories as nodes of an ontology in Ontology-Based Network Construction, which become candidate category nodes to be finally selected as document labels in Spreading Activation. Thanks to this extension, the ontology-based network can be applied to interest extraction for marketing communication. As for the second problem, we do not use inter-article links in Ontology-Based Network Construction since it is considered that they have a negative effect on spreading activation. By removing inter-article links, category nodes can be activated appropriately. Figure 3 shows the difference between the existing and our proposed method.

5.3 Methodology

In this subsection, we describe our extensions to Ontology-Based Network Construction step (step 1) . We also explain the differences between the existing and proposed methods for the internal implementation of Initial Node Selection step (step 2) .

First, the extensions to Ontology-Based Network Construction step are described as follows.

1-a. Construct the initial network: We construct an ontology-based network based on the link structure of Wikipedia using article-category links and inter-category links.

1-b. Add nodes for IAB categories: We add IAB categories as nodes in the network, manually select a Wikipedia category node that has the same or similar name as each IAB category node, and create links from the selected Wikipedia category nodes to the IAB category nodes.

Second, the internal implementation of Initial Node Selection step is described as follows.

2. Initial Node Selection: We use the profile text and the most 200 recent posts of each Twitter account as an input document. We extract the Wikipedia article titles in the input document, score these titles by tf-idf value, and select the article nodes corresponding to the top k articles as the initial nodes. tf-idf, short for term frequency–inverse document frequency, is often used as a weighting factor for words and documents in information retrieval and text mining. The existing method also uses tf-idf as a word weighting score. We set k to 40 because the average number of Wikipedia article titles in an input document is about 40.

To validate the effectiveness of our proposal for interest extraction, we conducted two preliminary evaluation experiments; manual evaluation and validation using data from company accounts. We show the results that validate the effectiveness in Appendix B of the extended version [16] due to the space limitation. The manual evaluation shows that the proposed extension improves the performance of interest extraction. Also, we validate that the extracted interests of accounts can be used to predict new followers.

6 FOLLOWER PREDICTION MODEL

In this section, we construct a follower prediction model for providing interpretable results that are useful in deciding the direction of marketing communication on social media. To this end, we utilize random forest[5] because it is a widely used classifier and its results are interpretable. By using both candidate account selection methods and interest vectors, the classifier calculates the probability of how likely each account follows a given company account. As for candidate account selection methods, we adopt follower- and similar-company-based selections because they perform higher than keyword-based selection (see Table 2). Thanks to this model, we can obtain predicted accounts that have high probability of following the company account. Also, we can identify which dimensions of interest vector largely affect the quality of the classifier by analyzing the learned model. The remainder of this section describes the procedure of building the follower prediction model.

Candidate Account Selection We first extract candidate account set V_{cand} by follower- or similar-company-based selections. We

[5]https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

denote the set of existing followers as $V_{existing} \subseteq V_{cand}$. Remember that we assume the recent followers are informative in predicting new followers for follower-based selection. Here, we denote time window size as s and candidate account set in the most recent s months as $V^s_{cand} \subset V$ for follower-based selection. For the common parts of follower- and similar-company-based selections, both V_{cand} and V^s_{cand} are referred as V_{cand} for simplicity.

Data Used for Training Our Model The attributes of account i are expressed as an interest vector, $x_i \in \mathbb{R}^{361}$. Note that there are 361 advertising categories provided by IAB Categorization and they are used as Twitter content categories. For training data, the existing followers $V_{existing}$ are used as positive samples, and accounts that do not follow a company account $V_{cand} \backslash V_{existing}$ are used as negative samples, where \ indicates set difference operation. Because the size of V_{cand} is generally much larger than that of $V_{existing}$, we perform random sampling for negative samples so that the number of negative samples becomes the same as that of the positive samples.

Classifier We construct a binary classifier that predicts whether accounts follow a given company account. We train this classifier using the training data described above. Then, the model computes the follower probability for the accounts in $V_{cand} \backslash V_{existing}$. Since we predict new followers from the accounts that do not still follow a given company account, we use negative samples for training data as test data.

7 EXPERIMENTS

The goal of this section is to validate the effectiveness of our proposal and the interpretability of its results through a case study. In Section 7.1, we give experimental settings and describe existing methods for comparison. In Section 7.2, we conduct extensive experiments to compare our proposal with various existing methods. Finally, we also identify the discriminative dimensions of interest vectors, which are important to predict new followers in Section 7.3.

7.1 Experimental Settings

7.1.1 Dataset. We collect and use data of the accounts of four companies, A, B, C, and D (see Table 1 for details). The size of V_{cand} varies depending on companies; therefore, we perform random sampling to make $|V_{cand}| = 3.0e5$. With regard to follower-based selection, we perform random sampling to make $|V^s_{cand}| = 3.0e5$, $6.0e5$, $9.0e5$, respectively for training data with time window size $s = 1, 2, 3$ (months)[6].

7.1.2 Comparisons for Feature Vector. To validate the effectiveness of our interest vectors, we compare them with two baseline embedding methods, SIF and word2vec, which are insufficiently interpretable. SIF [3] is widely used as a baseline for sentence embedding. word2vec [19] is a well-known word embedding method. We use the average of the embedding vectors for words that appear in the posts or profile as the feature vector of each account. We use the open word2vec model constructed from a large Web corpus including Twitter data [17]. Word embedding models can achieve high performance in 200 or 300 dimensions for many tasks [18]. We chose 200 as the dimensionality of the word2vec model, which is considered as a reasonable size. Due to the nature of the method, the dimension of the sentence embedding by SIF is 200, the same as the dimension of the word embedding used. We first extract two intermediate vectors $x^{prof}, x^{post} \in \mathbb{R}^{200}$ from the profile and posts for each account, respectively. Then, we create feature vector $x^{emb} \in \mathbb{R}^{400}$ of the account by concatenating x^{prof} with x^{post}.

[6]This is because the size of candidate account set increases with time window size s.

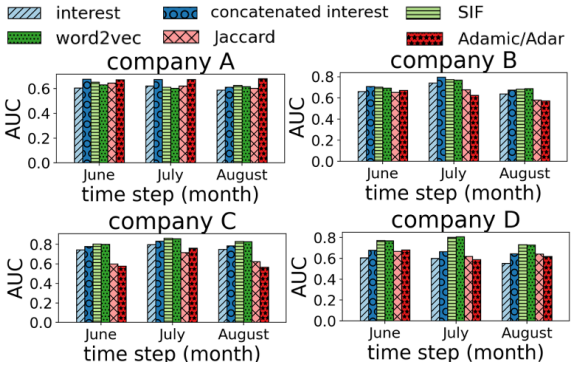

Figure 4: AUC of Follower Prediction Using Follower-based Selection for the Month of τ.

7.1.3 Comparisons for Link Prediction. Most link prediction methods cannot be applicable because they need to additionally collect links between accounts in V_{cand} as mentioned in Section 3.1. Exceptionally, few simple methods can be applicable, such as Jaccard index and Adamic/Adar since they utilize only common neighbors, which are collected by follower-based selection. We implement link prediction methods based on these metrics. Intuitively, these metrics measure how many common neighbors each candidate account shares with a given company. In our experiments, we directly use Jaccard index and Adaminc/Adar for link prediction. Also, we concatenate them with our interest vector and use the vector, called **concatenated interest**, as a feature vector for prediction.

In summary, we compare two types of our interest vectors (interest and concatenated interest) with several existing methods as follows. We train random forest models with either interest, concatenated interest, SIF, or word2vec. Also, we calculate Jaccard index and Adamic/- Adar for each account in V_{cand} and directly use them as the probability of the account.

7.2 Prediction Performance Comparison

We conduct experiments using various time steps and time window sizes to evaluate the effectiveness of our proposal.

7.2.1 Evaluation Measure. To assess the results, we adopt the area under the receiver operating characteristic (AUC) that is a commonly used measure to evaluate the predictive ability of classifiers.[7] We report the mean of AUC scores of five runs. Note that we do not adopt precision, recall, or F1 to compare prediction results, since those scores are sensitive to additional threshold parameters, so they make the comparison difficult.

[7]We also report evaluation results by using another measurement, the area under the precision-recall curve (AUPR), in Appendix C.1 of the extended version [16] due to the space limitation. The results validate that our interest vectors obtain promising results in most cases.

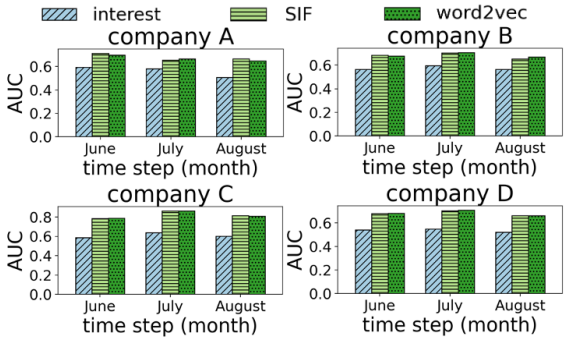

Figure 5: AUC of Follower Prediction Using Similar-Company-based Selection for the Month of τ.

7.2.2 Variation in Time Step. We predict new followers at month of τ, June, July, or August, using the candidate account sets obtained by either follower-based selection or similar-company-based selection. As for follower-based selection, we set the time window size s to 1 month since one month of data can be easily collected. We will investigate how time window size affects to the prediction results in Section 7.2.3. Note that we can calculate link prediction metrics, Jaccard and Adamic/Adar, for follower-based selection since the selection collects common neighbors shared by a given company and candidate accounts.

First, we observe that the models using our interest vectors obtain promising results for link prediction. Figure 4 shows the results using follower-based selection by changing month of τ. The results demonstrate that the models using the concatenated interest obtained competitive AUC scores to the models using the embedding methods (SIF and word2vec) for companies A and B and obtained better scores than simple link prediction methods (Jaccard, Adamic/Adar) in most cases. We also observe that the models using the embedding methods outperform the models using our interest vectors in many cases. Figure 5 shows the results using similar-company-based selection by changing month of τ. The results demonstrate that the models using the embedding methods obtained higher AUC scores than our interest vectors. The score difference between our interest vectors and the embedding methods implies the trade-off between interpretability of the feature vectors and prediction performance.[8]

Next, we observe that the models using follower-based selection achieve higher AUC scores than those using similar-company-based selection for all feature vectors (interest, SIF, and word2vec), which are applicable to the two selections. Table 3 shows the average AUC scores obtained by the models using either of follower- or similar-company-based selections. The models using follower-based selection outperform those using similar-company-based selection in most cases. We conjecture that the reason is that the followers of similar companies tend to have similar interests in general, so it makes difficult to learn binary classifiers. As shown in Table 2, follower- and similar-company-based selections are comparable in terms of precision ratios of new followers to candidate accounts. Hence, we conclude that follower-based selection outperforms similar-company-based selection for overall performance.

[8]To further analyze the trade-off, we conducted experiments for a non-interpretable classifier, multi-layer perception (MLP). We show the detailed results in Appendix C.2 of our extended version [16]. We obtained an interesting observation that there is at most only 4% difference between AUC scores of random forest and MLP using embedding vectors.

Table 3: Average AUC Scores Obtained by Models Using Either of Follower-based Selection or Similar-Company-based Selection for Feature Vectors, Interest, SIF, and Word2vec. Bold numbers indicate the best results.

Company	Selection method	June	July	August
Company A	follower	0.6291	0.6117	**0.6107**
	similar company	**0.6678**	**0.6339**	0.6072
Company B	follower	**0.6858**	**0.7611**	**0.6693**
	similar company	0.6416	0.6685	0.6281
Company C	follower	**0.7816**	**0.8393**	**0.8028**
	similar company	0.7200	0.7883	0.7428
Company D	follower	**0.7152**	**0.7377**	**0.6725**
	similar company	0.6346	0.6544	0.6154

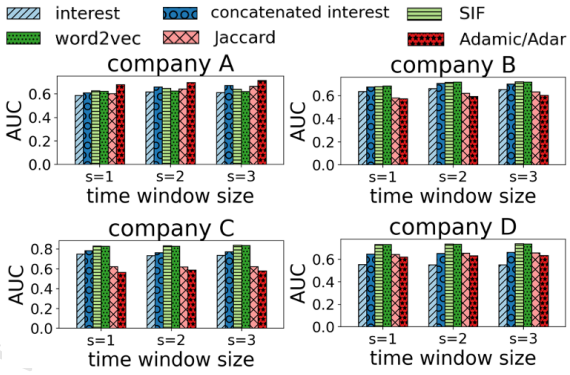

Figure 6: AUC of Follower Prediction Using Follower-based Selection for New Followers in August 2020 Using Cumulative Followers Acquired in Most Recent s Months.

7.2.3 Variation in Time Window Size. In this experiment, we investigate how time window size affects to the prediction results for follower-based selection. We do not evaluate the case for similar-company-based selection, since the followers of similar companies are not largely affected by the time window size. We conduct experiments for various time window sizes, $s = 1, 2, 3$. That is, we predict new followers in August 2020 using the followers acquired in the most recent s months.

Figure 6 shows the results of follower prediction using follower-based selection by changing time window size of s. We observe that AUC scores of all methods for each company are stable over various time window sizes. Hence, the models using our interest vector obtain promising results as discussed in Section 7.2.2.

7.3 Case Study

We demonstrate the interpretability of the results of our proposal. We set s to 1 month and predicted new followers in August 2020 by applying random forest with interest vectors. We use follower-based selection since the model using follower-based selection achieves better prediction results than using similar-company-based selection. To explain the interpretation of interest vectors, we show that

Table 4: Top 5 Dimensions of Interest Vectors in Feature Importance of Learned Random Forest. In parentheses, we show industry domains of companies.

	Company A (Fashion)	Company B (Food)	Company C (Cosmetic)	Company D (Restaurant)
1	Science/Physics	Business/Marketing	Business/Advertising	Business/Marketing
2	Science/Space, Astronomy	Business/Advertising	Beauty/Makeup, Cosmetic	Food&Drink/American Cuisine
3	Business/Marketing	Food&Drink/Japanese Cuisine	Business/Marketing	Beauty/Makeup, Cosmetic
4	Book&Literature/Health, Mind, Body	Book&Literature/Health, Mind, Body	Food&Drink/Japanese Cuisine	Food&Drink/Japanese Cuisine
5	Business/Advertising	Science/Space, Astronomy	Book&Literature/Health, Mind, Body	Beauty/Skin Care

our proposal clarifies which dimensions (factors) of the interest vectors contribute most largely to the prediction results.

Table 4 shows the top five dimensions in the interest vectors that contribute most largely to the prediction results by random forest. For companies B, C, and D, there are categories closely related to their industries; "Food&Drink/Japanese Cuisine" appears for company B that belongs to food industry, "Beauty/Makeup, Cosmeti" appears for company C that belongs to cosmetic industry, and "Food&Drink/American Cuisine" and "Food&Drink/Japanese Cuisine" appear for company D that belongs to restaurant industry. We conjecture that the reason why we could not observe categories related to the industry of company A is that it is difficult to predict new followers of company A as the AUC scores are lower than those of other companies (see the rows of follower-based selection in Table 3). We additionally observe that "Business/Marketing" is appeared in all company accounts. We conjecture that the followers of company accounts tend to have this kind of general category.

8 CONCLUSION AND LIMITATIONS

We addressed the follower prediction problem on social media, which identifies a set of accounts that are likely to follow a given company account. We first proposed several heuristic account selection methods, which can be performed with few social network API calls. We clarified that follower- and similar-company-based selection approaches perform with high precision. Next, we proposed an interest extraction method that extracts interpretable feature vectors (interest vectors) from accounts on social media by using standardized categorization. Finally, we constructed a follower prediction model utilizing the above two account selection methods and interest vectors. Through our experiments of real-world data, we demonstrated that our follower prediction model obtains promising results in various settings. We also demonstrated a case study of the interest categories related to companies for follower prediction.

Limitations Finally, we discuss the limitations of our proposal. We chose Twitter for our target application because the abundant text information in Twitter can improve the quality of follower prediction. However, some accounts rarely post anything, so the proposed method cannot be applied to those accounts. Also, our proposal is not so effective for social media that include less texts, such as Instagram. We conducted experiments by using only four companies drawn from a wide range of industries. While the number of companies is small and limited, the results gained demonstrate that our proposal clarifies which dimensions of our interest vectors contribute to follower prediction.

REFERENCES

[1] Lada A Adamic and Eytan Adar. 2003. Friends and neighbors on the web. *Social networks* (2003).

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of KDD*.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR*.

[4] Yixiang Fang, Xin Huang, Lu Qin, Ying Zhang, Wenjie Zhang, Reynold Cheng, and Xuemin Lin. 2020. A survey of community search over big graphs. *The VLDB Journal* (2020).

[5] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint* (2014).

[6] Li He, Yan Jia, Weihong Han, and Ding. Zhaoyun. 2014. Mining user interest in microblogs with a user-topic model. *China Communications* (2014).

[7] IAB Technology Laboratory. 2020. CONTENT TAXONOMY. https://iabtechlab.com/standards/content-taxonomy/.

[8] Leslie K John, Oliver Emrich, Sunil Gupta, and Michael I Norton. 2017. Does "liking" lead to loving? The impact of joining a brand's social network on marketing outcomes. *Journal of Marketing Research* (2017).

[9] Jaeyong Kang and Lee. Hyunju. 2017. Modeling user interest in social media using news media and wikipedia. *Information Systems* (2017).

[10] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of ICML*.

[11] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML*.

[12] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. 2007. The Dynamics of Viral Marketing. *ACM Transaction on the Web* (2007).

[13] Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL (Short Papers)*.

[14] Kwan Hui Lim and Amitava Datta. 2013. Interest Classification of Twitter Users Using Wikipedia. In *Proceedings of WikiSym'13 + OpenSym'13*.

[15] Qing Liu, Minjun Zhao, Xin Huang, Jianliang Xu, and Yunjun Gao. 2020. Truss-based community search over large directed graphs. In *Proceedings of SIGMOD*.

[16] Seiji Maekawa, Santi Saeyor, Takeshi Sakaki, and Makoto Onizuka. 2021. Extended Version: Effective Candidate Selection and Interpretable Interest Extraction for Follower Prediction on Social Media. *https://github.com/seijimaekawa/Follower_Prediction_Extended* (2021).

[17] Shogo Matsuno, Sakee Mizuki, and Takeshi Sakaki. 2019. Constructing of the word embedding model by Japanese large scale SNS + Web corpus. In *Proceedings of the Annual Conference of JSAI2019*.

[18] Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The Role of Context Types and Dimensionality in Learning Word Embeddings. In *Proceedings of NAACL*.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

[20] Kazuki Nakajima and Kazuyuki Shudo. 2020. Estimating Properties of Social Networks via Random Walk considering Private Nodes. In *Proceedings of SIGKDD*.

[21] Feng Qiu and Junghoo Cho. 2006. Automatic Identification of User Interest for Personalized Search. In *Proceedings of WWW*.

[22] Basit Shahzad, Ikramullah Lali, M. Saqib Nawaz, Waqar Aslam, Raza Mustafa, and Atif Mashkoor. 2017. Discovery and classification of user interests on social media. *Information Discovery and Delivery* (2017).

[23] J David Smith and My T Thai. 2020. Measuring Edge Sparsity on Large Social Networks. In *Proceedings of ICWSM*.

[24] Zareen Saba Syed, Tim. Finin, and Anupam Joshi. 2008. Wikipedia as an Ontology for Describing Documents. In *Proceedings of ICWSM*. The AAAI Press.

[25] Twitter. 2020. Ad Targeting Best Practices for Twitter. Twitter Business. *https://business.twitter.com/en/targeting.html* (2020).

[26] Twitter. 2021. More about restricted uses of the Twitter APIs. *https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases* (2021).

[27] Marijke De Veirman, Veroline Cauberghe, and Liselot Hudders. 2017. Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude. *International Journal of Advertising* (2017).

[28] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint* (2017).

[29] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. 2015. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* (2015).
[30] Ryen W White, Peter Bailey, and Liwei Chen. 2009. Predicting user interests from contextual information. In *Proceedings of SIGIR*.
[31] Yang Yang, Ryan N Lichtenwalter, and Nitesh V Chawla. 2015. Evaluating link prediction methods. *Knowledge and Information Systems* (2015).
[32] Fattane Zarrinkalam, Mohsen Kahani, and Ebrahim Bagheri. 2018. Mining user interests over active topics on social networks. *Information Processing & Management* (2018).
[33] Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. In *Proceedings of NeurIPS*.

A VALIDATION OF TIME SERIES EFFECT

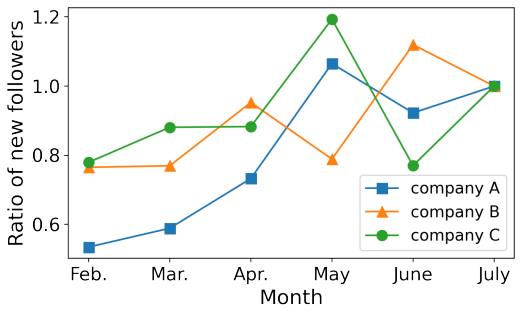

Figure 7: Precision Ratios of Follower-Neighbor Sets. We regularize the ratios by the ratio of the latest month (July).

We validate our assumption that recent followers are more informative to predict new followers than other existing followers. The motivation is that the API limitation makes it difficult to collect follow-relationships for all followers of a company account. We investigate the precision ratio of candidate account sets extracted by follower-based selection during the last six months in order to show the effectiveness of the recency of following time. We utilize the followers of companies A, B, and C[9] and treat the followers acquired in August 2020 as new followers. Figure 7 shows the precision ratio of the candidate account set for each month. We remove the effect of the sizes of candidate account sets by dividing their sizes with the number of new followers because the number of followers varies from month to month. In the figure, we observe that the candidate account sets of more recent months include new followers with higher precision in both companies.

Practically, advertisers can easily collect follow-relationships of their recently acquired followers. On the other hand, it is impossible to find out when the followers followed a company caused by the limited function of the social network API. For example, if advertisers want to collect data for a month, they can begin effective follower-based selection after a month. This observation shows that the time series effect contributes to follower-based selection.

Table 5: Target 22 Categories for Manual Evaluation

Japanese Cuisine	Baseball	Spas
Celebrity Fan/Gossip	Investing	Advertising
Law, Gov't & Politics	Beauty	Tennis
Family & Parenting	Movies	Television
Comic books	Marketing	Fashion
Health & Fitness	Theme parks	Football
Computer Games	Golf	Animation
Computer Networking		

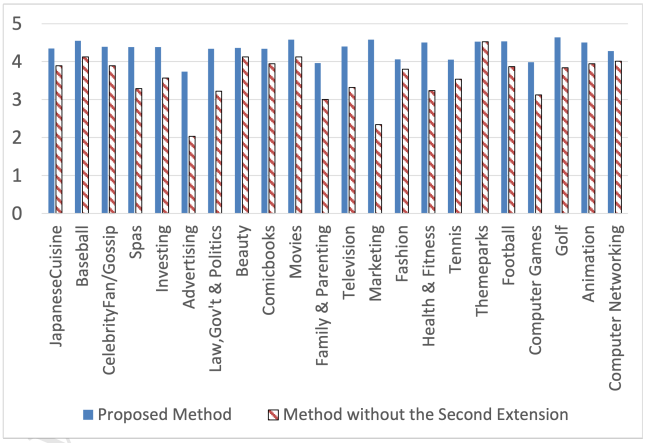

Figure 8: Results of Manual Evaluation of Interest Estimation.

B PRELIMINARY EVALUATION EXPERIMENT FOR INTEREST EXTRACTION

B.1 Manual Evaluation

Since we cannot know true interest categories of user accounts, we conduct manual evaluation for measuring how appropriate the interest vectors of accounts are to the accounts. Also, we compare the method with a method without the second extension that removes inter-article links to appropriately activate category nodes, in order to validate that the second extension improves the performance of the proposed method. Note that the first extension is required to apply methods to interest extraction for marketing communication as mentioned in Section 5.2.

We extract the interests of one million accounts, rank the categories in the order of the number of extracted accounts, and choose the top 22 IAB subcategories as the evaluation targets, which are presented in Table 5. We evaluate the validity of the 100 accounts that belong to these 22 categories. In this experiment, we set the number of iterations for the activity propagation to $5(t_0 \leq t \leq t_5)$. For each category, the profile and keywords of the accounts for evaluation are presented to the three evaluators, who answer the relevance of the category on a five-point scale from 1 (no information indicating interest in the category) to 5 (the category is the main interest).

[9]We choose companies A, B, and C because their followers can be collected more easily than those of company D in terms of the number of followers (see Table 1).

The average evaluation scores for each category are shown in Figure 8. This figure shows that the proposed method achieved 4.0 or higher scores for 19 categories, except for the categories of "Advertising," "Family & Parenting," and "Computer Games". The average score for all 22 categories was 4.34. This indicates that the results are highly convincing to the evaluators in many categories.

We also show the result of the method without the second extension in Figure 8. We observe that the proposed method outperforms the method without the second extension for most categories. This indicates that the second extension effectively improves the quality of interest extraction of the proposed method.

B.2 Interest Validation Using Data from Company Accounts.

We investigate whether the interest vectors extracted by the proposed method can be effectively used to predict new followers of a given company account. We aim to demonstrate that the interest vectors of new and existing followers are similar, when compared with the non-following accounts. Therefore, we examine the similarity between the interest vectors among the sets of new followers, existing followers, and non-following accounts in a candidate account set for each company account. In this experiment, we use follower-based selection. Let τ be the month for which we predict new followers; these sets can be represented as follows:

- new : A set of new followers in the month of τ
- $existing$: A set of existing followers before the month of $\tau - 1$
- not : A candidate account set in the month of $\tau - 1 \setminus$ a set of followers ($new \bigcup existing$), where $\setminus$ denotes the set difference operation.

Note that these three sets are subsets of the candidate account set in the month of $\tau - 1$.

The verification procedure is as follows. First, we create an interest vector for each account and normalize the sum of the vectors for each account to 1. Next, we calculate the average of each dimension of the vectors for new followers, existing followers, and non-following accounts. In this verification, we conducted the experiment using three periods of $\tau = 6, 7, 8$ for companies A, B, C, and D. We adopt the Kullback-Leibler (KL) divergence, a widely used metric to calculate the difference in distributions.

The evaluation results are listed in Table 6. The table shows that in most cases, the KL of the interest vectors for new and existing followers ($new \mid existing$) is smaller than that of the interests for new followers and non-following accounts ($new \mid not$). Through this verification, we show that there is a difference between the interest vectors of followers and non-following accounts, even in a candidate account set.

C ADDITIONAL EXPERIMENTS

C.1 Evaluating Prediction Results on AUPR

To assess the results of follower prediction by multiple measures[10], we adopt the area under the precision-recall curve (AUPR) as the measure, which is suitable for imbalanced data in terms of the numbers of positive and negative samples [31]. We report the mean of the AUPR scores of five runs. We focus on follower-based selection that outperforms similar-company-based selection in Table 3. In the experiments, we set the time window size s to 1 month and use random forest as in Section 7.2.2.

[10]Note that we report the follower prediction results by AUC in Section 7.2

Table 6: Kullback-Leibler Divergence (KL) of Interest Vectors between New Followers and Existing Followers and between New Followers and Non-Followers. KL indicates the difference between two input distributions.

		June	July	August
company A	$new \mid existing$	0.217	0.130	0.157
	$new \mid not$	0.199	0.153	0.178
company B	$new \mid existing$	0.088	0.050	0.075
	$new \mid not$	0.125	0.225	0.086
company C	$new \mid existing$	0.054	0.028	0.025
	$new \mid not$	0.209	0.355	0.233
company D	$new \mid existing$	0.485	0.148	0.121
	$new \mid not$	0.316	0.174	0.115

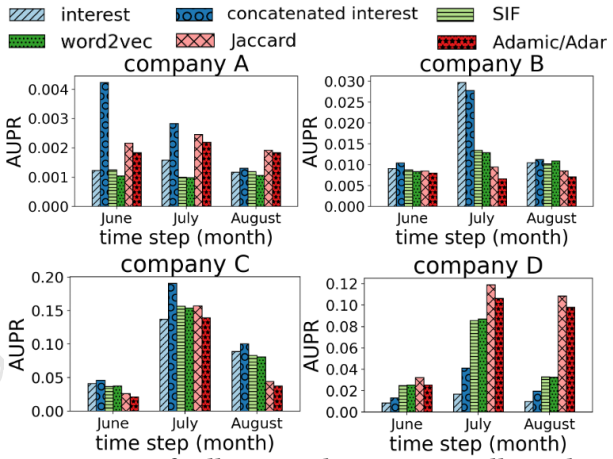

Figure 9: AUPR of Follower Prediction Using Follower-based Selection for New Followers in the Month of τ.

Figure 9 shows the results of AUPR for follower-based selection by changing month of τ. We observe that the models using our interest vector obtain promising results for follower prediction. Consequently, the models using our interest vectors obtain promising results with both AUC and AUPR.

We observe that the AUPR scores are relatively low for all time steps and all companies. This is mainly because the used data is highly imbalanced between positive and negative samples since the number of followers is largely smaller than the number of candidate accounts. Actually, Yang, Lichtenwalter, and Chawla [31] reports the same imbalanced class distributions in the link prediction problem and also their AUPR scores are comparable to ours.

Also, we report that the AUPR scores are not stable for time steps. This is because AUPR is sensitive to positive samples and few positive samples largely affect the AUPR scores.

C.2 Non-interpretable Classifier

To further analyze the trade-off between interpretability of the feature vectors and prediction performance, we conducted the

Table 7: Performance Comparison of Random Forest (RF) and Multi-Layer Perceptron (MLP) Using Embedding Vectors. Gain represents the performance improvement of MLP from RF.

			June	July	August
Company A	SIF	RF	0.651	0.612	0.627
		MLP	0.661	0.631	0.641
		Gain	1.46%	3.02%	2.20%
	word2vec	RF	0.631	0.602	0.617
		MLP	0.648	0.625	0.640
		Gain	2.62%	3.68%	3.56%
Company B	SIF	RF	0.703	0.774	0.682
		MLP	0.714	0.789	0.702
		Gain	1.53%	1.85%	2.90%
	word2vec	RF	0.694	0.768	0.688
		MLP	0.701	0.781	0.691
		Gain	0.91%	1.75%	0.47%
Company C	SIF	RF	0.802	0.863	0.831
		MLP	0.812	0.869	0.832
		Gain	1.29%	0.67%	0.08%
	word2vec	RF	0.800	0.858	0.827
		MLP	0.814	0.864	0.828
		Gain	1.70%	0.68%	0.06%
Company D	SIF	RF	0.772	0.802	0.733
		MLP	0.731	0.819	0.737
		Gain	-5.58%	1.98%	0.62%
	word2vec	RF	0.768	0.810	0.731
		MLP	0.792	0.822	0.730
		Gain	2.94%	1.53%	-0.11%

experiments for multi-layer perceptron (MLP) that is a typical non-interpretable classifier and usually achieves high classification results. We focus on follower-based selection that outperforms similar-company-based selection in Table 3. In the experiments, we set the time window size s to 1 month as in Section 7.2.2. To avoid bias, the best set of the hyperparameters of MLP was determined using the Optuna software [2].

Table 7 shows the AUC scores of random forest (RF) and MLP using embedding vectors, SIF and word2vec which obtain high AUC scores in Figure 4. The result indicates that MLP achieves higher AUC scores than RF in most cases (see Gain in Table 7). We also observe that the performance gains are at most 3.68%. Through these experiments, we observe that MLP obtains the relatively small improvement of prediction performance against random forest if we sacrifice the interpretability of a classifier.