#FewThingsAboutIdioms: Understanding Idioms and its Users in the Twitter Online Social Network

Koustav Rudra[1], Abhijnan Chakraborty[1], Manav Sethi[1], Shreyasi Das[1], Niloy Ganguly[1], and Saptarshi Ghosh[2,3]

[1] Department of CSE, Indian Institute of Technology Kharagpur, India
[2] Max Planck Institute for Software Systems, Germany
[3] Department of CST, Indian Institute of Engineering Science and Technology Shibpur, India

Abstract. To help users find popular topics of discussion, Twitter periodically publishes 'trending topics' (trends) which are the most discussed keywords (e.g., hashtags) at a certain point of time. Inspection of the trends over several months reveals that while most of the trends are related to events in the off-line world, such as popular television shows, sports events, or emerging technologies, a significant fraction are *not* related to any topic / event in the off-line world. Such trends are usually known as *idioms*, examples being #4WordsBeforeBreakup, #10thingsIHateAboutYou etc. We perform the first systematic measurement study on Twitter idioms. We find that tweets related to a particular idiom normally do not cluster around any particular topic or event. There are a set of users in Twitter who predominantly discuss idioms – common, not-so-popular, but active users who mostly use Twitter as a conversational platform – as opposed to other users who primarily discuss topical contents. The implication of these findings is that within a single online social network, activities of users may have very different semantics; thus, tasks like community detection and recommendation may not be accomplished perfectly using a single universal algorithm. Specifically, we run two (link-based and content-based) algorithms for community detection on the Twitter social network, and show that idiom oriented users get clustered better in one while topical users in the other. Finally, we build a novel service which shows trending idioms and recommends idiom users to follow.

1 Introduction

Twitter is now considered more of an 'information network' than a social network [6] and almost the entire focus of the research community has been on 'topical' content in Twitter, such as tweets / hashtags related to sports or technology or emergency situations in the off-line world [2]. However, a closer inspection of the Twitter *trending topics* ('trends' in short) – keywords periodically declared by Twitter as being the most discussed at that point in time – indicates some exceptions to this view, and provides the motivation for the present study.

We collected US trends over a duration of 10 months (January – October, 2014) using the Twitter API at 15-minute intervals. This gave about 18,500 distinct trending topics during this period. We then developed a classifier *Odin*[4] and classified the trends

[4] Named after the God of Wisdom according to Norse mythology; details in Section 2.

Category	%	Example trends
Entertainment	33%	#5sosonKiis, #IWishICould, #Austinonidol
Sports	30%	#argentinavsholanda, #lakers, #bravsger
Idioms	9%	**#WhenIWasATeenager, #FactsaboutMe, I get angry when**
Technology	8%	#iphone6, #galaxy4, AppleWatch, ios8
Politics	5%	#tcot, #pjnet, #obama, #gaza
Business	5%	#amazon, #AlibabaIPO, #FedReserve
Religion	3%	#EidMubarak, #jesus, #citrt
Health	2%	#Ebola, #Who, #breastcancer
Others	5%	#garlicparmpizza, #filipino, cheesecake, pizza is healthy

Table 1: **Percentage of Twitter trends collected over ten months, and classified into nine different categories that were identified by human volunteers (details in Section 2). Also given are few examples of trends.**

into multiple categories such as sports, entertainment, technology etc. – these broad categories were identified by human volunteers (details in Section 2). Table 1 shows the distribution of the trends in the different broad categories. While most of the categories are topical and related to events in the off-line world, it can be observed that a special category, known as *idioms*[5], regularly becomes trending. Examples of idioms include #4WordsBeforeBreakup, #11ThingsAboutYou, and apparently these are not related to any topic or event in the off-line world.

The frequent presence of such trends is intriguing – it raises the question whether their dynamics as well as the users discussing such trends are similar to those of the topical counterparts. To understand the dynamics, we collected tweets related to hundreds of idioms and the users who discuss them, and conducted a detailed measurement study. We find that the tweets containing idioms are mainly conversational in nature; for instance, they hardly contain URLs. On investigating the users who post the tweets (the idiom-users), we find that they are mostly general and active Twitter users, as opposed to being popular experts / celebrities who usually drive topics such as politics and entertainment. The idiom-users maintain close friendships among themselves and interact on diverse issues with their friends. Thus, the study unfurls that hidden within the *information network* of Twitter, there is a *social network* of users who regularly have "non-topical" conversations among themselves.

Such an inference has far-reaching implications. It essentially means that multiple dominant dynamics are present in the same social network – so the standard tasks like community detection, recommendation, and so on, *cannot* be done using a one-parameter-fits-all approach. An algorithm to identify (recommend) topical groups might fail to identify (recommend) idiom-users. To test this proposition, we run two community detection algorithms – one identifying topical groups [2] and the other, Infomap [13] which detects communities using link structure. We find that the idiom-users are well identified by Infomap while the topical groups are better identified by [2]. This establishes that different approaches for tasks such as clustering may have different utilities in a heterogeneous online social network. Further, considering that all existing

[5] In this study, we follow the definition of idioms given by [12] – an idiom is a keyword representing a conversational theme on Twitter, consisting of a concatenation of at least two common words which does not include names of people, places or music albums etc.

recommender services are specifically meant to recommend topical experts, we develop a service *Idiomatic* where one can easily follow popular idiom-users, see the recent and past trending idioms and post tweets using them.

2 Classification of Trends

In order to perform a large scale study on idioms and the trends related to topics / events in the off-line world, we built an automatic classifier *Odin*, to distinguish particular trends based on whether they are idioms or related to some topic. Note that some prior studies [7, 20] have also attempted to classify trends (not necessarily into the same categories found by Odin), utilizing the textual contents of the tweets containing the trends. However, tweets (restricted to 140 characters) often contain informal language and abbreviations which potentially results in lower classification accuracy [20]. Hence, we adopt a different approach that combines both tweets and related web documents and uses several web-based knowledge engines to perform the classification. Odin classifies a given trend following the steps presented below.

2.1 Preprocessing

Segmentation: Trends often consist of multiple words [12] recognizing which is easy for multi-word phrases and hashtags written in CamelCase style (e.g., #WorldCupSoccer), but is very difficult for trends which simply have the words juxtaposed without any separation (e.g., #everythingididntsay). Since it is important to identify the individual words which make up a trend in order to understand its topic, trends need to be segmented into the component words. Odin follows a modified version of the Viterbi Algorithm [1], which uses a model of word distribution to calculate the most probable character sequence forming a word. Odin computes the word distribution from Google n-gram corpus (*https://books.google.com/ngrams*). Given a trend, Odin segments the trend into its constituent words based on this calculated probability estimates (details omitted for brevity).

Categorization of related web documents: Odin searches different Web search engines (e.g., Google, Bing) with the segmented trend, to get a large set of web-pages relevant to the given trend. Often the tweets containing the trend have URLs, which become another source for getting related web-pages.[6] For a given trend, Odin collects all the web-pages pointed from the tweets and returned by the search engines; and then a set of category keywords are extracted for these collected web-pages using the NLP-based AlchemyAPI web service (`www.alchemyapi.com`).

Entity extraction and categorization: Sometimes the trend contains names of people, organisations or locations (e.g., #EMABiggestFansJustinBieber) detecting which might give a clear idea on the category of the trend. Similarly, the web documents and the tweets associated with a particular trend have many such named entities present in them. Odin extracts such entities using AlchemyAPI and then queries Freebase (`www.freebase.com`) to know the 'notable type' of such named entities (e.g., according to Freebase, notable type for 'Justin Bieber' is '/music/artist').

[6] URLs leading to social media sites like Facebook, Twitter, Instagram, are ignored, since these pages usually do not have much content to help the topic categorization process.

Property	Idiom	Sports	Entertainment	Technology
Number of trends	150	150	150	150
Total #tweets containing the trends (millions)	6.205	6.787	6.967	6.105
Mean #tweets per trend	41,369	45,257	46,455	40,721
Total #distinct users posting the trend (millions)	2.74	2.71	1.90	1.75
Mean #distinct users per trend	18,315	18,098	12,725	11,705

Table 2: **Statistics of data collected**

2.2 Classification

At the end of preprocessing steps, for a given trend, Odin collects the categories of the related web documents and the notable types of the related named entities. Treating the number of web documents and named entities in the various categories as features, Odin uses a Support Vector Machine (SVM) classifier with Radial Basis Function kernel to classify a particular trend into one of the 9 categories shown in Table 1.

Training Data Preparation: For creation of training data, three human volunteers (regular users of Twitter, who are not authors of this paper) were asked to manually inspect 700 distinct trends collected during the first two weeks of January 2014 (along with tweets containing these trends), and classify the trends into different categories. The volunteers identified the *nine broad categories* shown in Table 1, such as Entertainment, Sports, Technology, Idioms (following the definition of idioms in [12]). Out of the 700 trends, all three volunteers agreed upon a particular category for 575 trends. We created the training data considering this unanimous categorization as the ground truth.

Classification Performance: Standard 10-fold cross validation on the data of the 575 trends showed that Odin attains 77.15% accuracy in predicting trend categories, which is good considering that it is a complex nine-class classification task.

3 Dataset

Since most of the Twitter trends were related to the three topics *entertainment*, *sports*, and *technology* (see Table 1), we decided to focus on idioms and trends related to these three topics; the trends related to any of these three topics are collectively referred to as '*topical trends*'. For each of the trends belonging to the four selected categories, we collected as many tweets containing the trend as possible using the Twitter search API. To get a better understanding about the trends, in our analysis as presented in later sections, we used only those trends for which we were able to collect more than 30,000 tweets. To maintain uniformity across categories, we finally selected a set of 150 trends related to each of the categories (the actual distribution is stated in Table 1).

For each of the 600 selected trends, we further collected detailed statistics about all the users (including their profile details, social links and recently posted tweets) who posted a tweet containing any of the selected trends. Table 2 summarizes the statistics of the data collected for the trends of the four categories.

4 Comparing Idioms and Topical trends

In this section, we compare how idioms and topical trends are discussed in the Twitter social network, and the users who discuss them frequently.

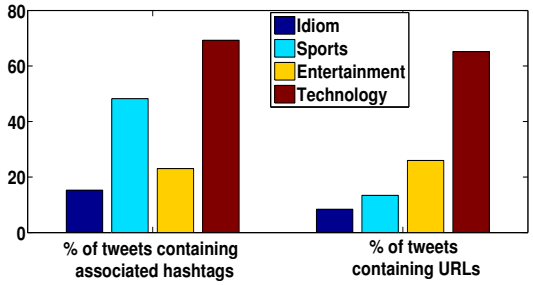


Fig. 1: **Comparing topical trends and idioms: Percentage of tweets which contain (i) other hashtags (apart from the trend under consideration), and (ii) URLs. All values averaged over all trends of a particular category.**

4.1 How trends are discussed in Twitter

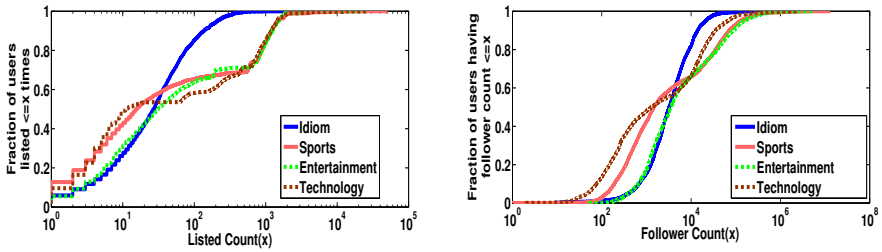
We first analyze how the trends of different categories are, in general, discussed in Twitter. For a given trend t, we consider all tweets containing t, and measure what percentage of these tweets contain other hashtags (apart from t itself), and URLs.

Figure 1 shows mean values of the percentage of tweets containing other hashtags and URLs, where the mean values are computed over all trends of a particular category. Statistical measures like two sample KS-test and Mann-Whitney U test with significance level 0.05 show that there is a significant difference in the distribution of the mean values among the four categories. Expectedly, we find that the topical trends are much more likely to be accompanied by other hashtags and URLs related to the corresponding event in the off-line world. For instance, the sports-related trend #LIVvCHE (referring a match between the two English soccer clubs Liverpool and Chelsea) is accompanied by the hashtag #Torres which indicates a player who is a part of the match. On the other hand, Twitter-specific idioms are very seldom accompanied by other hashtags since they are not related to external websites or news-stories in the off-line world.

We also observed the timeline evolution of trends, i.e., how they start getting tweeted and become popular in Twitter. Expectedly, most topical trends emerge as a result of some related event in the off-line world, such as a sports or musical event, or a socio-political incident / issue. In case of idioms, an interesting pattern observed is that many idioms *initially* propagate along with hashtags related to some specific event in the off-line world. For example, the idiom '#MyFavouriteActor' first appeared with the hashtag '#PeoplesChoice' (related to the People's Choice awards), while the idiom '#SexRequirements' initially appeared with the health-related hashtag '#FitnessPromo'. These idioms, however, follow their independent path with users innovating interesting comments and thus making them popular.

4.2 Characterising users interested in various categories

In order to understand the nature of the users who are interested in promoting particular types of trends, we identify sets of users who are interested in the different categories (sports / technology / entertainment / idioms), and compare various characteristics of these users.



(a) **CDF listed count** (b) **CDF follower count**

Fig. 2: **Distribution of Listed & Follower count of four categories (idiom, sports, entertainment, technology) of users**

Identifying users interested in a certain category: To identify users who are interested in a certain category, we identify those users who *frequently* discuss trends of that category. For a particular category, we initially consider all the users who have posted at least one tweet on a trend in that category. We rank the users based on the number of different trends in that category on which they have posted at least one tweet. Subsequently, for each category, the 10,000 users who have posted most number of distinct trends in that category (according to our dataset) are considered. Since our objective is to identify users who are genuinely interested in trends of a certain category, we next attempt to verify whether the users selected above frequently discuss trends on that category. For this, we collected the 3,200 most recent tweets for each of the selected users, by crawling their time-line through the Twitter API, and used our classifier *Odin* to classify the hashtags contained in these tweets, to check what fraction of these hashtags were related to that category. For instance, for a certain user u included among the top 10,000 users who posted on most sports-related trends in our dataset, we checked whether a significant fraction of all hashtags included in u's recent tweets were related to sports. Additionally, Opencalais (**`www.opencalais.com`**) tool is used to identify the topic of each tweet present in the timeline of a user. We included a user in the final selected set for a category, if at least 30% of the hashtags and 70% of the tweets posted by her (among her recent tweets) were judged to be related to that category.

In this way, we finally identified a set of 5,000 users who are genuinely interested in each of the four categories. We refer to these sets of users as *idiom-users*, *sports-users*, *entertainment-users*, and *technology-users*. The rest of this section studies the characteristics of these sets of users.

Popularity of the users: We start by checking the popularity of the users interested in the various categories. We use two standard metrics of popularity of users in the Twitter social network [3, 4] – (i) follower-count, i.e., the number of followers of a given user, and (ii) listed-count, i.e, the number of Twitter Lists a given user is included in.[7] Both metrics resulted in very similar observations. Figure 2 shows the distribution of the listed-count and follower-count values of users who predominantly discussed the trends in the four categories.

[7] Lists are a feature by which a user can group together accounts on a common theme [4, 15].

Idiom-users			Sports-users			Entertainment-users			Technology-users		
Bio	**Lists**	**Tweets**	**Bio**	**Lists**	**Tweets**	**Bio**	**Lists**	**Tweets**	**Bio**	**Lists**	**Tweets**
life	faves	friend	sports	wwe	game	5sos	band	show	news	social	iphone
love	ily	people	football	wrestling	season	justin	album	music	tech	media	ios
fun	luke	hobby	wrestling	sports	team	bands	music	video	tech	tech	android
cool	nigg	niall	wwe	chelsea	nfl	ariana	youtubers	photo	oracle	tweet	google
harry	styles	school	soccer	cricket	football	luke	idols	album	software	business	apple

Table 3: **Characterizing the users who frequently discuss trends in each category – top 5 words appearing in (i) the user-bio in the profile, (ii) Lists in which the users are members, and (iii) tweets posted by the users.**

We observe an interesting trend. Almost *all* idiom-users are relatively less popular – 65% of the idiom-users have listed-count values in the range 0–40. In contrast, a significant fraction of the users who predominantly discuss the topical trends(sports, entertainment, technology) are very popular users, which includes experts from their respective fields. The above statistics lead to some interesting insights. There seems to be two very distinct types of users who dominantly participate in discussions on topics related to the off-line world (e.g., sports, entertainment, technology) – (i) popular users who are experts on these topics (e.g., researchers, sports-persons, journalists, musicians), and (ii) the common masses who are interested in these topics. This agrees with findings in recent research studies [2, 19]. In sharp contrast, users who dominantly participates in idioms are mostly common masses.

How the users are described: We next focus on how the users who are interested in various categories describe themselves, and how they are described by others. To infer the characteristics of a given user u, we refer to two sources – (i) the bio of u, which is a short description written by the user to describe herself, and (ii) the name and description of Twitter Lists in which u is included as a member – this indicates how other users (those who created the Lists and added u as a member) describe u [15, 17]. For a given category, we consider the bio (or List names and descriptions) of all the 5000 users chosen for this category (as described above), and find the words which occur in the bio (or Lists) of most number of these users.[8]

Table 3 shows the top 5 words which appear in the bio and Lists of the users for each category. As expected, the users for the topical categories (sports, entertainment, technology) are characterized by words related to the topics. For instance, sports-users are described by 'wrestling', 'wwe', entertainment-users are identified by '5sos', 'justin', and technology-users by 'social', 'tech'. On the other hand, the idiom-users are mostly described by words related to day-to-day conversation and positive sentimental words such as 'love', 'life', 'faves', 'ily (i love you)' and so on.

Content posted by the users: We next focus on the content (tweets) posted by the users. Similar to the previous analysis, we consider the set of tweets posted by the users interested in a certain category, and find out the most frequent words in the tweets. Table 3 shows the top 5 words posted by users in each category – we again find that while

[8] The bio and List-names are pre-processed using standard techniques such as case-folding, removal of a common set of stopwords, and so on.

User-group	Mention Network		Subscription Network	
	Reciprocity	**Density**	**Reciprocity**	**Density**
Idiom	21.88%	0.0012	49.57%	0.0221
Sports	14.67%	0.0017	10.19%	0.0030
Entertainment	13.40%	0.0010	13.76%	0.0058
Technology	13.91%	0.0011	4.87%	0.0025

Table 4: **Reciprocity and density of the mention and subscription networks among different groups of users.**

the sports-users, technology-users and entertainment-users mostly post words related to the corresponding topics, the idiom-users mostly use conversational words and phrases related to musical events, celebrities etc.

4.3 Studying the interactions among the users

We now investigate how the users in the four groups interact among themselves. In Twitter, the primary ways by which a user u can interact with another user v are (i) u can subscribe to the content posted by v by following v, or by following a List which has v as its member, and (ii) u can @mention v in her tweet.

Analysing interaction networks: We construct two types of interaction networks among the users. The first is a *subscription network* where a directed link $u \rightarrow v$ indicates that user (node) u subscribes to the content posted by user v. The second is a *mention network* where the link $u \rightarrow v$ indicates that user u has @mentioned v.

To quantify the level of interaction among the users, we measure two structural properties of the subscription and mention network – (i) *density*, which measures what fraction of all links which can be present in a network, are actually present, and (ii) *reciprocity*, which indicates what fraction of the directed links are reciprocated, i.e., both the links $u \rightarrow v$ and $v \rightarrow u$ exist in the network. The importance of reciprocity is that if two users share a reciprocal link, then the two users are *mutual friends* with a higher probability (as compared to the chance of a fan subscribing to a celebrity, but the celebrity not reciprocating).

Table 4 shows the reciprocity and density of the mention and subscription networks among different groups of users. We find that the density of the subscription network among the idiom-users is significantly higher compared to that for the sports-users, entertainment-users, and technology-users. Also, the reciprocity is significantly higher for both the subscription network and the mention network for idiom-users, indicating that a large fraction of the interactions are between mutual friends. These observations indicate that, just like users interested in a common topic (sports, entertainment or technology), the idiom-users form their own group; in fact, they subscribe to / mention one another much more frequently than the topical groups of users.

Note that the density of the mention networks are comparable for all the user-groups. This is because, as observed earlier, the sports-users, technology-users, and entertainment-users contain a large number of common (less popular) users and a few popular celebrities, and most of the @mentions result from the common users mentioning the celebrities. For instance, a signficant fraction of the @mentions among

User-group	Idioms	Sports	Entertainment	Technology
% of topical hashtags in retweeted tweets	22.83%	78.47%	81.63%	79.57%
% of topical hashtags in mentioned tweets	25.74%	74.58%	77.12%	78.54%

Table 5: **Percentage of hashtags (present in tweets) where a user of a certain group mentions or retweets another user of the same group, which are related to the topic of interest of that user-group.**

technology-users are directed towards @twitter and a few software companies. However, the reciprocities are lower for the topical groups, since the celebrities do *not* mention the common users. On the other hand, most of the mentions among the idiom-users (who have similar popularity) come from conversations among mutual friends, leading to high reciprocity. In fact, as much as 62.5% of the mentions among the idiom-users are between two users who share a *reciprocal* link in the corresponding subscription network (i.e., are likely to be mutual friends), where as this percentage is less than 35% for the topical user-groups.

Nature of conversations among the users: Finally, we analyze the nature of the conversations among the users of the same group. Specifically, when a user retweets or mentions another user *in the same group*, we check whether the hashtags used in the tweets are related to the common topic of interest of the users. For instance, among the hashtags which a sports-user retweets or mentions to another sports-user along with the tweets, we check what fraction of such hashtags are related to sports. For this, we use our classifier *Odin* to classify hashtags present in the tweets where a user mentions or retweets another user from the same-group. The results are shown in Table 5. More than 74% of the hashtags that are mentioned / retweeted among the sports-users, entertainment-users, and technology-users are related to the corresponding common topic of interest of that user-group. In sharp contrast, only about 25% of the hashtags that are exchanged among idiom-users are idioms. This again shows that idiom-users are not a focused topical group rather they engage themselves in diverse issues.

4.4 Type of user-groups and their identifiability

Our analyses reveal that the group of users interested in Twitter-specific idioms has very different characteristics compared to the groups of users interested in topics such as technology, sports and entertainment. In this section, we attempt to explain the differences and their implications on identifiability of the groups.

Explaining group formation: Formation of user-groups in a social network has been a long-standing topic of research in sociology, and several theories have been proposed to explain their formation [8, 18]. According to the well-accepted *common identity and common bond theory* [5,10,11], there are two primary types of groups. In *identity-based groups*, people join the group due to their interest in a well-defined common theme (topic), whereas *bond-based groups* are driven by personal social relations (bonds) among the members, and may be characterized by the absence of any common topic of discussion. As a result, bond-based groups have higher reciprocity among the members than identity-based groups. Also, the discussions in bond-based groups tend to vary widely and cover multiple subjects, while in identity-based groups, they tend to be related to the common topic of interest of the group.

User-group	Idioms	Sports	Entertainment	Technology
Nos. communities	107	284	272	281
Nos. groups assigned per user	9	2	2	3

Table 6: **(i) Number of communities identified by Infomap, into which a user-group is scattered, (ii) average number of topical groups assigned per user by the topical group identification approach developed in [2].**

The above analyses on the four user-groups show that, as expected, the users interested in a common topic like sports, entertainment or technology form identity-based groups, with fewer interactions (@mentions) among friends, and most of the discussions among the members being related to the topic of common interest (Table 5). On the other hand, the idiom-users group is characterised by relatively higher levels of personal interactions with mutual friends, and a relatively small fraction of the conversation among the friends is related to their common topic i.e. idioms. Hence, the idiom-users form a bond-based social community within Twitter, in which they discuss their personal topics of interest as well as conversational matters.

Identifiability of the groups: The differences in the nature of various user-groups can have significant impact on the identifiability of the groups. To demonstrate this, we used two algorithms for detecting groups in the Twitter social network, and checked how well they could identify the idiom-users group and the topical groups.
(i) We used the well-known Infomap community detection algorithm [13] on the Twitter subscription network among all the users spanning the four user-groups. Then we enumerated the number of different communities identified by Infomap, where the members in any of the four user-groups are distributed. Table 6 (second row) shows that the topical groups were scattered into significantly higher number of Infomap communities, as compared to the idiom-users group.
(ii) Bhattacharya et al. [2] proposed a methodology to identify *topical communities* in Twitter (comprising of users who are experts on a topic or interested in the topic). We used this method to check the number of distinct topical communities a member in our dataset is placed. We found that, on average, a user in the idiom-users group is placed in many more topical communities, than a user in the sports-users / entertainment-users and technology-users groups (Table 6, last row).
These observations reveal that within Twitter, there exist two different kinds of network structure – one is an information network, and other one is social communication network. Any community detection method which considers only one facet of the network might not be able to identify all the communities accurately.

5 Idiomatic: Service for Idiom Lovers

As stated earlier, the focus of the research community has been entirely on the topical content discussed in Twitter, such as identifying experts on various topics [4, 16]. However, for a user who is interested in idioms (idiom lover), there is no existing service to recommend whom she could follow to know interesting idioms being discussed in Twitter. Hence, we have developed *Idiomatic* (**`http://cse.iitkgp.ac.in/resgrp/cnerg/idiomatic`**), a service where one can easily follow popular idiom-users (ranked according to the number of idioms they post), have a quick look at recent

and past trending idioms (classified by an enhanced version of the Odin classifier presented in Section 2 from continuous stream of trending topics collected at 15 minute intervals), and post tweets using idioms.

To evaluate the quality of the recommended idiom-users, we used human feedback since relevance of user-profiles to a certain topic / theme is subjective in nature. The evaluators were shown the most recent 100 tweets of the idiom-users, and were asked to judge whether the user appears to be an active idiom-user or not. 15 human volunteers individually judged the top 20 idiom-users shown by the service. Out of the top 20 users, 18 were judged as active idiom-user by *all* the evaluators, and even the remaining two users were judged as active idiom-users by majority of the evaluators.

6 Related Work

The present study focuses on the characteristics of Twitter idioms, identifying users who actively participate in idioms, and understanding the social behaviour of the groups of these users. Some prior studies on trending topics in Twitter have focused on classification of the trends [9, 20], whereby the presence of idioms [20] is identified. However, there has been little effort in analyzing the characteristics of idioms, and of the users who post the idioms. To our knowledge, the only prior study which attempted to compare idioms with trends related to events in the off-line world is by Naaman et al. [9], where they used different features like content, interaction etc. to classify the trends. However, they did not attempt to analyze the users who discuss such idioms.

Also note that there have been prior attempts to distinguish between bond-based and identity-based groups in online social networks (see Section 4.4). For instance, [14] classified chats among users on a text-based communication platform into two categories – on-topic chats which are on a common topic (identity-based) and off-topic chats where people chatted on a variety of topics (bond-based). More recently, [2] identified a large number of topical groups in Twitter, comprising of users who are experts or seekers of information on various topics, and showed that these groups are essentially identity-based. In this work, we explored the nature of the groups among the idiom-users, and found that they reveal bond-based characteristics.

7 Conclusion

The popular perception of the research community is that, there are two parts of Twitter – one interesting part where participants read and post a wide variety of topical tweets, and another part which comprises of pointless babble and is hence unimportant and uninteresting. However, in our study, we find that these pointless babbles, even though not related to any off-line event, frequently become trending in Twitter due to participation of large number of common masses. These users form bond-based groups among themselves to discuss their personal interests – idioms and some other forms of fun and gossip. This study has several implications, e.g., for community detection in social networks. Keeping in mind the popularity of idioms, we developed a whom-to-follow recommendation service where idiom lovers can easily find trending idioms and users who post idioms actively and frequently.

References

1. Berardi, G., Esuli, A., Marcheggiani, D., Sebastiani, F.: ISTI@TREC Microblog Track 2011: Exploring the Use of Hashtag Segmentation and Text Quality Ranking. In: NIST TREC (2011)
2. Bhattacharya, P., Ghosh, S., Kulshrestha, J., Mondal, M., Zafar, M.B., Ganguly, N., Gummadi, K.P.: Deep Twitter Diving: Exploring Topical Groups in Microblogs at Scale. In: ACM CSCW (2014)
3. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: Proc. AAAI ICWSM (May 2010)
4. Ghosh, S., Sharma, N., Benevenuto, F., Ganguly, N., Gummadi, K.: Cognos: crowdsourcing search for topic experts in microblogs. In: Proc. ACM SIGIR (2012)
5. Grabowicz, P.A., Aiello, L.M., Eguiluz, V.M., Jaimes, A.: Distinguishing topical and social groups based on common identity and bond theory. In: Proc. ACM WSDM (2013)
6. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proc. World Wide Web Conference (WWW) (2010)
7. Lee, K., Palsetia, D., Narayanan, R., Patwary, M.M.A., Agrawal, A., Choudhary, A.: Twitter Trending Topic Classification. In: Proc. IEEE International Conference on Data Mining Workshops (2011)
8. McMillan, D., Chavis, D.: Sense of community: A definition and theory. Journal of Community Psychology 14(1), 6–23 (1986)
9. Naaman, M., Becker, H., Gravano, L.: Hip and trendy: Characterizing emerging trends on Twitter. Journal of the American Society for Information Science and Technology 62(5), 902–918 (2011)
10. Prentice, D.A., Miller, D.T., Lightdale, J.R.: Asymmetries in attachments to groups and to their members: Distinguishing between common-identity and common-bond groups. Personality and Social Psychology Bulletin 20(5), 484–493 (1994)
11. Ren, Y., Kraut, R., Kiesler, S.: Applying Common Identity and Bond Theory to Design of Online Communities. Organization Studies 28(3), 377–408 (2007)
12. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: Proc. World Wide Web Conference (WWW). pp. 695–704 (2011)
13. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. PNAS 105, 1118–1123 (2008)
14. Sassenberg, K.: Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. Group Dynamics Theory Research And Practice 6(1), 27–37 (2002)
15. Sharma, N., Ghosh, S., Benevenuto, F., Ganguly, N., Gummadi, K.: Inferring Who-is-Who in the Twitter Social Network. In: Proc. WOSN Workshop (2012)
16. Twitter Help Center — Twitter's suggestions for who to follow, https://support.twitter.com/articles/227220-twitter-s-suggestions-for-who-to-follow
17. Wagner, C., Liao, V., Pirolli, P., Nelson, L., Strohmaier, M.: It's not in their tweets: Modeling topical expertise of twitter users. In: Proc. ASE/IEEE SocialCom (2012)
18. Wenger, E.: Communities of practice: Learning, meaning, and identity. Cambridge University Press, New York, NY, USA (1999)
19. Wu, S., Hofman, J.M., Mason, W.A., Watts, D.J.: Who says what to whom on Twitter. In: Proc. World Wide Web Conference (WWW) (2011)
20. Zubiaga, A., Spina, D., Fresno, V., Martínez, R.: Real-Time Classification of Twitter Trends. Journal of the American Society for Information Science and Technology (2014)