2015 IEEE 15th International Conference on Data Mining Workshops

# Exploiting Class Bias for Discovery of Topical Experts in Social Media

Iuliia Chepurna and Masoud Makrehchi
Department of Electrical, Computer, and Software Engineering
University of Ontario Institute of Technology, Oshawa, ON, Canada
Email: {iuliia.chepurna,masoud.makrehchi}@uoit.ca

***Abstract*—Discovering contexts of user's expertise can be a challenging task, especially if there is no explicit attribution provided. With more professionals adopting social networks as a mean of communicating with their colleagues and broadcasting updates on the area of their competence, it is crucial to detect such individuals automatically. This would not only allow for better follower recommendation, but would also help to mine valuable insights and emerging signals in different communities. We posit that topical groups have their unique semantic signatures. Hence, we can treat identification of expert's topical attribution as a binary classification task, exploiting the class bias to generate training sample without any manual labor. In this work, we present profile- and behavior-based models to explore experts topicality. While the former focuses on the static profile of user activity, the latter takes into account consistency and dynamics of a topic in user feed. We also propose a naïve baseline tailored to a domain used in evaluation. All models are assessed on a case study of Twitter investment community.**

## I. Introduction

Rapid adoption of online social platforms has drastically changed the landscape of what was perceived as traditional media. New communication paradigm empowers literally anyone to broadcast their message to millions. Breaking news witnessed by locals, reaction towards major events and controversial social issues, perception of brands and political leaders, professional advice, even details of daily routines—all that is constantly shared through social networks and microblogs. Acting as an outlet, they provide a unique opportunity to gaining recognition to these highly dedicated users who devote their time to crafting content of extreme value. And while the latter is curated by a small group of elite individuals, the masses of ordinary users rely on them for disseminating interesting information. Latest survey shows that roughly half of participated Facebook and Twitter users regularly consume news on these platforms [1]. Moreover, content producers themselves actively exploit this medium: 54% of US journalists report to find their stories on Twitter, and 79% monitor social media for breaking news [2].

However, absence of content verification on such systems makes them vulnerable to spread of rumors and misinformation. Hence, the task of identifying individuals authoring credible and engaging material is of utmost importance. While opinions of famous users are often in the spotlight, we are interested in discovering individuals that are *experts* in particular fields. Since the notion of knowledge exists only within a specific context (at most couple of them), we aim to detect *topical experts* as opposed to influential users that are famous across different communities.

While some platforms allow users to explicitly sign up for groups establishing communities based on their interests, others do not support such functionality. For example, to overcome this issue on Twitter people form implicit network by following alike [3]: politicians subscribe to politicians, journalists listen to journalists, entrepreneurs track other successful peers, and so on. Automatic detection of such groups has a number of applications, such as professional follower recommendation, extraction of up-to-date trends in specific domains, finding reliable business intelligence sources, surveillance of suspicious activities, and others.

Although most of the research effort in expertise localization has focused on generating small list of the most influential people with respect to the field of their knowledge, mostly to be consumed as a source for a follower recommendation, our objective is different. Motivated by a number of works successfully utilizing experts' recommendations for decision making in financial domain [4]–[7], we are interested in obtaining a representative sample of users whose collective intelligence can be later applied in external prediction task. Albeit the financial prediction is not of our primary interest in current work, keeping this application in mind, we would like selected pool of professionals to be as much diverse in terms of the level of their expertise as possible [8]. That is in this context detecting individuals with average or even marginal expertise is equally important to identifying most knowledgeable users.

We propose an automatic approach to topical community detection in social media platforms, such as Twitter, based on user authored content. We define two models to capture both user's pertinent and temporary interest in studied domain. Our approach takes advantage of relatively small number of professionals in the area of interest as opposed to the whole population of the platform in order to generate training set, this way eliminating manual construction of the ground truth. We also design a naïve baseline which exploits background knowledge about the domain, and does not require any training. We evaluate these models on Twitter's community of professional investors.

The rest of the paper is organized as follows. We review previous work on identification of expert's topicality and

978-1-4673-8493-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDMW.2015.198

knowledge discovery in section II. All proposed models as well as specifics of chosen domain are discussed in section III. We present description of the datasets and evaluation scenarios in IV, and summarize the results and directions of future work in VI.

## II. Previous Work

Significant research effort has been undertaken with respect to topical experts identification. Very first works were concerned with experts retrieval from knowledge databases which were manually curated within organizations [9], [10]. Since then focus of the research has shifted to mining of existing documents authored by potential candidates, this way allowing to detect qualified individuals without a need in constructing skill databases.

Number of diverse media channels has been explored for this purpose: starting with enterprise corpora [11], [12], followed by discussion groups [13] and Q&A communities [14], with most of research attention concentrated lately on various social media. As opposed to enterprise emails and document repositories, content generated on online social platforms is publicly available, thus making this source extremely attractive for research community. Besides, different types of user interactions supported by these systems provide a possibility to tap into collective opinion of studied community and judge which individuals are perceived as knowledgeable and which are not.

Previous works define experts within a context of one or several areas of their proficiency, thus differentiating between topicality and expertise. While some studies touch upon topical relevance—either it is treated simply as a query matching task [12], [13], [15], [16] or tackled by topics modeling [3], [11], [17], [18]—most of the research endeavor has focused on users' expertise. Various approaches have been proposed, which can be roughly divided into *network-based* and *feature-based*. The former [3], [13], [19]–[21] are variations of PageRank that tend to propagate leaders' influence through explicit or implicit networks. They exercise the intuition that users interacting with authorities would have higher influence than those without such a tie. However, such approaches normally favor general authorities and may easily overlook newly joined users, and they are also computationally expensive. The feature-based techniques have explored many dimensions of candidate representation, covering various aspects of user-attributed content [15], [17], [18], [22], [22], [23], structure of his static and dynamic networks [18], [19], [22], [23], engagement with the platform and patterns of his temporal activities [14], [15], [18], [22].

Other works have studied how quality of expert identification varies across different media. Guy *et. al.* [12] provided a comprehensive overview of enterprise social applications, such as blogs, wikis, bookmarking, file sharing and others, and showed that profile tags and microblogs demonstrate superior performance. It is expected, since descriptive tagging within an organization would normally boil down to major set of skills that colleagues believe candidate to posses. However, another cross-platform study on potential of LinkedIn, Facebook and Twitter in detecting knowledgeable individuals [15] showed quite surprising results. Not only LinkedIn was outperformed by Twitter, it showed the worst results among the channels. Moreover, Twitter has proven to be more informative than all three platforms combined. We believe that the cause of it lies in a short nature of Twitter updates (not more than 140 characters), which is expressive of dynamics and persistence of topics in candidate's conversations. It also supports user connections based on their affinity, yet without requiring these ties to be reciprocal. Based on discussed considerations, in this work we concentrate on the Twitter as our target platform.

A different line of research has actively explored this medium as well, particularly Twitter mechanism of list subscriptions [17], [19], [21], [24]. The latter allows users to group accounts they follow into some meaningful categories and provide them with descriptive annotations (*e.g.* finance, social computing, python development, *etc.*). The assumption is that community itself will discover most prominent individuals, if sufficient amount of users list them under the same or similar areas of expertise. However, despite the fact that working with Twitter lists yields significant results even for niche topics [19], such strategy also has its limitations. Namely, list-based approaches will fail if they need to discover experts that recently joined the network or motivated but novice individuals. Note that we do not want to bias our sample to the most distinguished practitioners, to the contrary, we would like this pool to express as much diverse opinions as possible.

Finally, we need to point out that there is a number of proprietary services for influence discovery, such as Klout[1], PeerIndex[2], Kred[3], Wefollow[4] and Twitter's own Who to Follow[5]. However, the problem with them is that, first, most solicit users to explicitly sign up to be discovered by the algorithm, and, second, the details of underlying implementation are not revealed to broad public.

Clearly, judging about individual's expertise is non-trivial, and actually very subjective task. For that reason most of the studies we discussed required vast amount of human participation either for evaluation of expert selection or generating ground truth. In this work, on the contrary, we are more interested in defining user's topical attribution rather than exact level of his expertise. We propose an automatic approach to topical community detection based on semantics prevailing within a group of interest, which is discussed in the next section.

## III. Proposed Approach

The community detection task is gaining more attention due to the recent growth of social media usage in a variety

[1] klout.com/corp/about (2015-05-06)

[2] peerindex.net/about.php (2015-05-06)

[3] kred.com (2015-05-06)

[4] wefollow.com/about (2015-05-06)

[5] support.twitter.com/articles/227220-about-twitter-s-suggestions-for-who-to-follow (2015-05-06)

of professional areas. Networks that were initially aimed for mundane communications, now are flooded with business transactions, targeted advertisement, dating opportunities, discussion boards for dedicated topics, and even illegal activities. One notable example is Twitter: its limitation on the post size forces users to produce more concise and informative content, this way making it a perfect medium for getting instant updates from various social circles user is involved in, including professional. However, discovery of such groups in systems which do not have an explicit mechanism of community membership can be a challenging task.

In this work we focus on identification of users' topical attribution based on the content they authored, and select Twitter as our target platform. We cast this problem as a binary classification of users with respect to domain of interest: a user can be relevant or irrelevant to selected domain. We hypothesize that each of these communities has a unique semantic signature: experts normally use shared lexicon, limited number of topics and even the same style of writing. Yet detection of such groups is not easy to generalize, since this task normally requires to have a background knowledge about the field.

Here, we present two models completely decoupled from the area of interest. They only require a positive example of the language used within a community, while the negative samples are generated automatically. Moreover, we do not impose any limitations on the medium positive samples originate from—it could be anything starting with news articles, blog posts, technical reports, interviews or even chapters from a textbook. We exercise the following intuition for obtaining negative training set: deciding whether a user belongs to specific community in a huge social network like Twitter is a typical example of classification in extremely imbalanced setting. With positive class being underrepresented, a probability of a randomly picked user to be considered as irrelevant is very high. Based on this premise Twitter streaming can be used to populate a negative set. Although there is an infinitesimal chance of a user being mislabeled, we believe our models to be robust to a small percentage of such errors in the training set. Henceforth we refer to the negative dataset as *white noise* and to positive—as *golden* set.

We devise two models—*profile-based* and *behavior-based*—to capture different aspects of user's engagement with the topic of selected domain. Former aggregates all tweets posted by a user in a single *profile* which is considered to be a representative proxy to user's interests. However, it is not able to differentiate between individuals who dedicate significant amount of their content to the topic of studied community and those rarely posting tweets relevant to the group. Also with *profile-based* model it is not clear how the decision should be made when considering a user who joined the group recently or the one who seem to lost interest in studied topic (a user who stopped posting relevant content). To address these issues we develop a *behavior-based* model which defines topical relevance on a tweet level. We also propose a baseline dependent on the field chosen for a case study. We briefly discuss the domain selected for our experiments and then provide a detailed description for each of the proposed models.

TABLE I
PART OF A SAMPLE TIMELINE OF A STOCK MARKET EXPERT FROM *target* TWITTER DATASET

| |
|---|
| From the daily chart, 625 in $AAPL wouldn't completely destroy the LT uptrend |
| In a few $SPY 145.5 weekly puts for 86c |
| Rainy mornings always throw my clock off for a bit |
| I'm up 103% and taking some cushion on the $SPY 145.5 puts I have |
| Drivin' along in my automobile.... http://t.co/HphiN1VS |
| A zoo. Kids request |
| So I guess the unlocked $YELP shares aren't immediately for sale |

**Selected domain.** Inspired by recent works that leverage experts' opinions for stock market prediction [4]–[7], we concentrate our attention on the investment community of Twitter. Obviously, most of the conversations revolve around performance of specific stocks, denoted by *cashtags*—ticker symbols preceded by a dollar sign (e.g. $AAPL, $MSFT, $TSLA, etc.). Oftentimes practitioners summarize their speculations with trading recommendations, such as BUY, HOLD or SELL. There is also a convention to finish stock-related tweets with a double dollar sign ($$), which is followed by many users. Note that since Twitter does not restrict users to sharing professional content only, those active practitioners would also have a significant number of posts related to personal matters. Actually, for half of the stock market experts on Twitter the fraction of professional content would not exceed the level of 0.33 (see Figure 2b). This potentially may lead to a confusion in our models, but we discuss this implication in more details in Experimental Results section.

We should discuss the choice of positive and target datasets as defined by our framework. As stated earlier, our goal is to identify topical groups on Twitter, thus *target* dataset is comprised of timelines of stock market experts active on Twitter. We describe how this dataset is obtained in the next section. Negative set, or *white noise*, is populated by selecting random users from Twitter stream. As can be seen from a sample timeline of such user (presented in Table II), their posts cover topics of general interest.

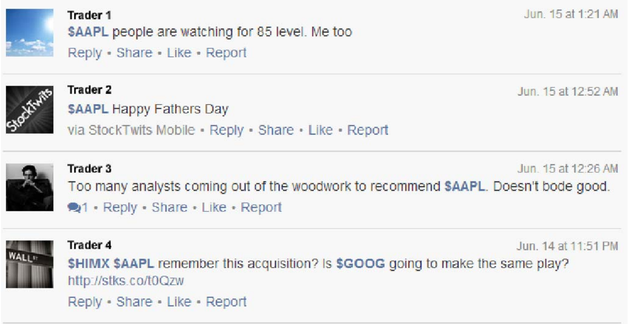
For the *golden*, or positive set, although there is a variety of sources to choose from, such as stock market discussion boards, individual blogs of selected advisors, news articles from financial data vendors (e.g. Bloomberg or Thomson Reuters, etc.) and many others, we restrict ourselves to *Stocktwits*[6]—microblogging service for investment community. The platform pretty much resembles Twitter with the main difference that users are solicited to share their insights on financial matters. Albeit there is no penalty or moderation of the content which is (slightly) off the topic (e.g. sometimes users wish each other a good weekend or productive week), experts prefer to use the medium for professional conversations only. Thus we believe that negligible fraction of irrelevant posts would be automatically discarded by the volume of financial content.

[6] stocktwits.com/about (2015-05-18)

You can see that majority of Stocktwits users have at lest 0.8 of their timelines considered to be relevant to the community. Similarly to Twitter, users of Stocktwits are limited to 140 characters to express their thoughts. Citation of a ticker by a cashtag is also strictly followed on this platform. For sample posts and user interface see Figure 1.

TABLE II
PART OF A SAMPLE TIMELINE FROM *White noise* DATASET

| |
|---|
| iPad 3 replacement screen from Apple was $299 ! So i just found a place on Yelp that does repairs and got it done for $80 ! |
| "We're the alchemists of our age, we don't know what we need to know yet" - @jjacoby #BICDCon |
| The dancing ghost loading animation in the Snapchat app is so amusing! http://t.co/x7WHjpsw7c |
| Martini's with Lisa (@ Holiday Inn Mart Plaza - Cityscape Bar) http://t.co/dZ84lOuwZc |
| Appreciate the little things in life |
| If you only knew... |



Trader 1 Jun. 15 at 1:21 AM
$AAPL people are watching for 85 level. Me too

Trader 2 Jun. 15 at 12:52 AM
$AAPL Happy Fathers Day
via StockTwits Mobile

Trader 3 Jun. 15 at 12:26 AM
Too many analysts coming out of the woodwork to recommend $AAPL. Doesn't bode good.

Trader 4 Jun. 14 at 11:51 PM
$HIMX $AAPL remember this acquisition? Is $GOOG going to make the same play?
http://stks.co/t0Qzw

Reply • Share • Like • Report
Reply • Share • Like • Report
1 • Reply • Share • Like • Report
Reply • Share • Like • Report
StockTwits
WALL ST

Fig. 1. Example of Stocktwits posts. Real usernames are replaced.

### A. Profile-Based

Let $T_i = \{t_1^{\langle i \rangle}, t_2^{\langle i \rangle}, \ldots, t_{N_i}^{\langle i \rangle}\}, N_i \in [1, N]$ denote a timeline of a user $i$, where $t_j^{\langle i \rangle} = \{w_1^{\langle i \rangle}, w_2^{\langle i \rangle}, \ldots, w_{M_{ij}}^{\langle i \rangle}\}, M_{ij} \in [1, M]$ is a tweet consisting of unigrams $w_k^{\langle i \rangle}$. Then *user profile* $U_i$ is represented by a binary vector space model

$$U_i = \left(w_1^{\langle i \rangle}, w_2^{\langle i \rangle}, \ldots, w_{|V|}^{\langle i \rangle}\right), \tag{1}$$

where $V = \bigcup_{i,j} w_{M_{ij}}^{\langle i \rangle}$ is a global vocabulary and

$$w_k^{\langle i \rangle} = \left\{ \begin{array}{l} 1, \text{ if } \exists\, j\colon w_k^{\langle i \rangle} \in t_j^{\langle i \rangle}, \\ 0, \text{ otherwise} \end{array} \right) \tag{2}$$

Each user $U_i$ is associated with a label $L_i \in \{0, 1\}$ with $L_i = 1$ indicating a user relevant to community. The *profile-based* model then uses samples $U$ and corresponding labels $L$ to fit a classifier.

In this model user is represented by a static *profile* compiled by aggregating all of his tweets into one document. We posit that such approach allow to detect salient topics[7] across user's lifetime on the platform. We restrict ourselves to unigrams instead of using more sophisticated language model, because, first, all three datasets—*golden*, *white noise* and *target*—exhibit different sets of topics (for *golden* they would be centered around performance of different types of equities, for *white noise* it would be a wide set of general topics, while the *target* would combine both), thus it is not valid to extract topic-words distributions jointly from these sets, and it might lead to a meaningless result; second, since all professional groups are known to use jargons, which usually consist of one word, the presence (or absence) of such lexicon could discriminate between relevant and irrelevant users. However, this model can be easily extended to higher order representations.

### B. Behavior-Based

In the *behavior-based* model user $U_i$ is represented as a collection of his individual tweets

$$U_i = \{w_{km}^{\langle i \rangle}\}, k \in [1, N_i], m \in [1, |V|], \tag{3}$$

where $w_{km}^{\langle i \rangle} = 1$, if $w_{km}^{\langle i \rangle}$ has occurred in $t_k^{\langle i \rangle}$.

To capture changes in behavior of the user with respect to topic of interest, *behavior-based* model builds the classifier for individual tweets as opposed to full timeline as in *profile-based* model. We also define vector of labels $l_i = (l_1^{\langle i \rangle}, l_2^{\langle i \rangle}, \ldots, l_{N_i}^{\langle i \rangle})$, where $l_k^{\langle i \rangle}$ is associated with each individual tweet $k$ for the given user $U_i$. For the training $l_k^{\langle i \rangle} = L_i$. To predict whether $U_i$ belongs to community or not, we calculate

$$r_i = \frac{1}{N_i} \sum_k l_k^{\hat{\langle i \rangle}} \tag{4}$$

where $r_i$ is the ratio of tweets considered as relevant by aforementioned classifier.

$$\hat{L_i} = \left\{ \begin{array}{l} 1, \text{ if } r_i > \theta, \\ 0, \text{ otherwise} \end{array} \right. \tag{5}$$

We would like to point that *behavior-based* model does not consider actual timing of the posts, which could have potentially shed some light on detecting users who just showed interest in the area and those who lost it long ago. However, it is able to reflect the notion of commitment—to what extent target user is dedicated to producing content related the field of interest as opposed to the one covering broad set of topics of a general interest. We expect this model to outperform *profile-based*, since it can reduce the rate of false positives, which are unavoidable in case of a static user representation. It is clear that it carries a significant computational overhead, since this method requires every single tweet to be classified by machine learning model.

### C. Domain-Specific Filter

To preserve the accuracy of *behavior-based* model while decreasing running time, we introduce a naïve baseline tailored for this specific domain. *Domain-specific filter* employs the same approach as in *behavior-based* with the only difference that instead of using a classifier it scans for occurrences of *cashtags* (e.g. $AAPL, $GOOG), and marks tweet as relevant

[7] Here and after in this section by the topic we mean topical (domain) orientation of the community to be detected.

if at least one was spotted. That is, if $C$ is a regular expression defining universum of cashtags, then we have

$$\hat{l}_k^{\langle i \rangle} = \begin{cases} 1, \text{ if } \exists\ m\colon w_{km}^{\langle i \rangle} \in C, \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

$\hat{L}_i$ is derived based on threshold $\theta$ similarly to *behavior-based* model.

The underlying idea is based on the fact that central object of discourse in this community is a specific company (or companies) represented by its ticker. Although presence of cashtags is not required for a tweet to be considered valid from investment standpoint, precise nature of these conversations results in every active expert citing a specific company at least once. Thus the only part left is to determine appropriate value for the threshold $\theta$. Also since the model does not require actual training, it is expected to have the fastest running time among the three.

We have to note that *domain-specific filter* cannot be easily generalized to all possible domains. However, one might come up with a slight modification: instead of a regular expression for cashtags one would have to generate a list of top unigrams from a lexicon used in target group and then check for their occurrences.

For this specific domain the goal is to determine whether a naïve baseline powered by a regular expression would suffice for a given task. If not, machine learning approaches introduced earlier should be used.

We also report that results of the baseline conform to our expectations regarding homogeneity of *golden* and *target* datasets (see Figure 2), with median of relevant posts equal to 0.8 for Stocktwits and 0.28 for Twitter.

## IV. Experimental Results

In this section we discuss the datasets collection, preprocessing procedures, experimental scenarios and the performance yielded by proposed models.

### A. Dataset description

**Target Twitter dataset** [4] was collected in four steps. First, using Twitter Search API we collected tweets for about 60 tickers based on manually generated list of company names and commonly used synonyms (e.g. “Apple Inc”, “AAPL”, “#AAPL” or “AAPL”). Then during March 27 till June 20, 2012, we streamed all tweets of users who authored first set of tweets returned by Search API. After that tweets of each individual user were automatically tagged as trading-related if: (i) tweet ended with double dollar sign (“$$”), or (ii) tweet contained at least one ticker and at least one of the predefined action words.[8] Users are considered to be traders only if they pass a threshold based on monthly, weekly and daily frequency of trading-related tweets. We ended up with 6512 users, out of which we randomly selected 1000 to constitute stock market experts dataset.

**Golden Stocktwits dataset** We randomly chose 1000 contributors and then crawled their timelines since the launch of Stocktwits (May 27, 2008) till March 31, 2014.

**White noise Twitter dataset** In order to collect representative timelines of randomly streamed Twitter users, we also employed two-step approach. First, we selected users authoring tweets collected from Twitter Streaming API, restricting ourselves to English tweets only. And then we collected historical data of every seed user. We limited *white noise* dataset to timelines of randomly chosen 1000 users. Please note that approach involving Twitter REST API (querying historical timelines) can yield up not more than 3200 of recent tweets per user. Statistics on the size of a timeline for each dataset are presented in Table III. As can be seen, with an average of more than 5K tweets per expert, *target* dataset is represented very well, while training datasets yield sufficient amount of data.

### B. Preprocessing

We performed a standard preprocessing procedures on the main datasets: we applied lower case, tokenized the documents into unigrams, removed stop-words, punctuation and digits. We also opted to ignore mentions (e.g. @twitter, @POTUS, etc.), hashtags (e.g. #RedNoseDay, #21demayo, etc.) and URLs. However, we decided to preserve all individual cashtags with preceding dollar sign being removed.

### C. Experimental scenarios

We tried binary (described in *Profile-Based* model) and inverse document frequency (idf) data representations with both of them yielding similar results. In terms of a classifier, we selected support vector machine (SVM) for both *profile-* and *behavior-based* models. Relevance thresholds were empirically set to $\theta_1 = \theta_2 = 0.3$. We further describe experimental settings used in number of scenarios.

**Scenario 1.1** is aimed to evaluate how good the model was fitted. Both supervised models (*profile-* and *behavior-based*) are trained on timelines of 1000 positive users from the *golden* set and 500 negative from the *white noise*. Performance is assessed using 10-fold cross-validation. Basic purpose of such setup is to ascertain that the models are capable of predicting unseen samples of the same type as trained with.

**Scenario 1.2** is of utmost interest for us, since it describes the expected *application setting*. Target users are predicted with model learned from automatically generated training set. To make sure that classifiers are not biased towards the positive

TABLE III
Statistics on the size of a user timeline for all used datasets.

| | Stocktwits | White Noise | Twitter |
|---|---|---|---|
| min | 1 | 1 | 3 |
| median | 6 | 69 | 3954 |
| avg | 173 | 264 | 5528 |
| std | 812 | 413 | 5857 |
| max | 13890 | 2800 | 48539 |

[8] Action words are defined as verbs occurring in a proximity to a ticker symbol with a high frequency across the whole dataset.

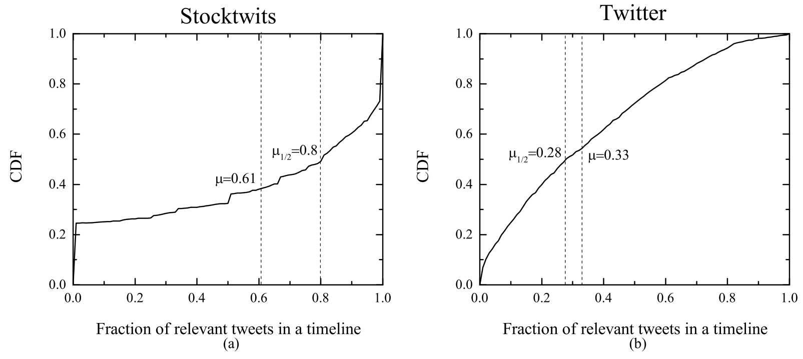


Fig. 2. Cumulative distribution function of a fraction of relevant tweets in timelines of (a) *golden* Stocktwits and (b) *target* Twitter datasets as determined by *domain-specific filter*. $\mu$ and $\mu_{1/2}$ denote mean and median of the distributions.

class, we add 500 negative users from *white noise* dataset to our target test set.

**Scenario 2.1** explores the validity of assumption that timelines of target Twitter users are noisy. This setup is similar to scenario 1.1 with the difference that *golden* set is now replaced with *target Twitter* dataset. Goodness of fit is tested using 10-fold cross-validation. We believe that *behavior-based* model will fail, because of confusion introduced by training tweets mistakenly tagged as positive. Recall that the fraction of tweets irrelevant to community (i.e. negative) is more than 0.67 for at least half of studied experts (see Figure 2). *Profile-based* model is believed not to degrade significantly. We elaborate discussion on results obtained for this scenario in the next subsection.

**Scenario 2.2** further explores the hypothesis of *noisy timelines* of target Twitter users. As we discussed before, majority of posts in Twitter timelines of selected experts is concerned with matters irrelevant to studied community. We believe that it might negatively affect the performance of application scenario. To validate the hypothesis, in this experiment we consider a setting exactly the opposite to the one described in scenario 1.2. That is *target* Twitter dataset is considered to be sufficient positive sample, and the models learned from it are tested on *golden* Stocktwits set. To make sure that none of the classes is discriminated, both training and test sets include negative users. We expect both machine learning models to fail in this experiment with *behavior-based* showing significantly worse performance.

*Domain-specific filter* is employed on the same testing scenarios with the difference that the actual training phase is omitted.

## V. Discussion

**Performance across designed scenarios:** We report the results obtained for 4 validation scenarios in Table IV. We list average, positive and negative $F$-measure, as well as accuracy, precision and recall.

Expectedly, all models including *domain-specific filter* performed well in cross-validation scenario when trained on clean data (scenario 1.1). Skewed class distribution (with number of negative samples equal roughly to half of positive) did not have any negative impact on either of experiments.

*Behavior-based* model showed even better performance ($F_1 = 0.96$ compared to 0.94) when tested on the target set (application scenario 1.2). Both *profile-based* model and *naïve filter* degraded significantly, yet achieving decent results ($F_1 = 0.78$ and $F_1 = 0.66$ respectively).

Interesting insights can be derived from results obtained in scenarios 2.1 and 2.2. As we expected, *behavior-based* model failed to cope with mislabeled training data: $F_1^- = 0.16$ and $recall = 0.54$ support the assumption on prevalence of false positives. Cross-validation scenario on *target* and *white noise* datasets did not affect performance of*profile-based* model, however, replacing test positive set by Stocktwits led to same consequence as for *behavior-based*.

We would also like to point out that in all scenarios where training data was not corrupted machine learning approaches have beaten the baseline.

**Performance in application scenario:** We now discuss how performance varies across different models for our aimed scenario 1.2 (see Figure 3). Although the *behavior-based* classifier has significantly outperformed other models due to its capacity to capture dynamics of topic usage by an expert ($F_1 = 0.96$), we cannot apply it in a real-time setting because of extreme computational overhead. *Naïve baseline* can be used for the tasks requiring testing of a vast amount of users—with near linear running time for a case of cashtag-based regular expression it achieves relatively decent performance of $F_1 = 0.66$.

For trivial scenarios which require higher accuracy, *profile-based* model seems to be the most suitable candidate: both time-wise and performance-wise it resembles a trade-off between first two.

**Dependency on the availability of posts:** One legitimate question stems from the fact that users are considered within

TABLE IV
COMPARISON OF THE RESULTS YIELDED BY PROPOSED MODELS. $F_1^+$ AND $F_1^-$ STAND FOR $F$-MEASURE OF POSITIVE AND NEGATIVE CLASSES RESPECTIVELY. PB DENOTES *profile-based* MODEL, BB—*behavior-based* MODEL AND DSF—*domain-specific filter*.

| | Model | Sc.1.1 | Sc.1.2 | Sc.2.1 | Sc.2.2 |
|---|---|---|---|---|---|
| $F_1$ | PB | 0.94 | 0.78 | 0.99 | **0.52** |
| | BB | 0.94 | **0.96** | **0.49** | **0.48** |
| | DSF | 0.81 | 0.66 | 0.70 | 0.81 |
| $F_1^+$ | PB | 0.96 | 0.81 | 0.99 | 0.45 |
| | BB | 0.96 | 0.97 | 0.81 | 0.81 |
| | DSF | 0.84 | 0.65 | 0.71 | 0.83 |
| $F_1^-$ | PB | 0.91 | 0.74 | 0.98 | 0.58 |
| | BB | 0.93 | 0.95 | **0.16** | **0.14** |
| | DSF | 0.78 | 0.66 | 0.69 | 0.78 |
| Precision | PB | 0.95 | 0.78 | 0.98 | 0.70 |
| | BB | 0.93 | 0.95 | 0.82 | 0.82 |
| | DSF | 0.82 | 0.75 | 0.76 | 0.82 |
| Recall | PB | 0.93 | 0.82 | 0.99 | 0.64 |
| | BB | 0.96 | 0.91 | **0.54** | **0.54** |
| | DSF | 0.72 | 0.74 | 0.77 | 0.86 |

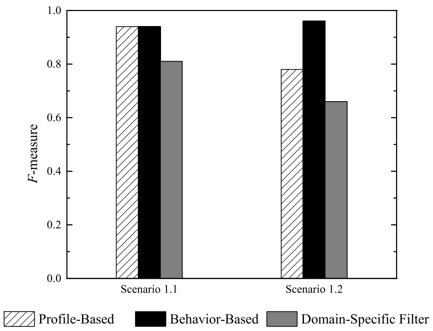


Fig. 3. Comparison of the results yielded by proposed models.

the context of their timelines. It is interesting to know how availability of their posts can affect the quality of prediction. Specially, considering the fact that content generation in the *target* dataset is described by power law (see Figure 4), even though all users are coming from the same homogeneous community, and data collection was not discriminating less active users.

Here we speculate how dependent devised models are on the timelines with a bigger size, or are they at all. We plot the prediction accuracy of target expert users (negative users are ignored in this setting) on Figure 5. Surprisingly, none of our models seem to rely on the size of user timeline available for testing. That is all models predict the relevance of users with around only 40 available tweets with the same accuracy as those with more than 5K tweets. It leads to a significant implication indicating that all of these models can be successfully used even on candidates with extremely small portion of a timeline observed, which means that even for those users who produce tremendous amount of information, we only have to analyze its small fraction to make a correct judgment.

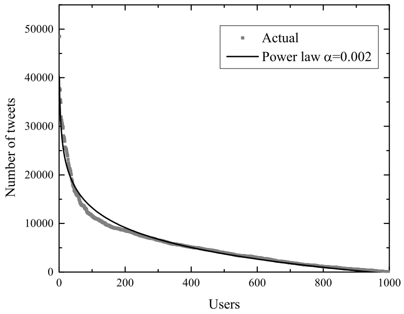


Fig. 4. Distribution of content generation in *target* Twitter dataset.

## VI. CONCLUSION

In this work we proposed an automatic approach to discovery of expert's topical attribution in social networks which do not allow its users to form explicit groups based on their interests. Presented approach exploits user authored content as a proxy to their interest. We casted this problem as a binary text classification task, and exercised the intuition that people within the same community would share the same semantic patterns. We described the procedure for automatic acquisition of training data based on the concept of extremely imbalanced binary classification. Our models require only a positive sample of a language used in the domain of interest with no restriction on the source this data is coming from. Negative examples are created automatically by randomly streaming Twitter feed. This way, unlike many other works, our framework does not need human participation (neither for annotating of the training set or for evaluating results). We devised two machine learning models—*profile-based* and *behavior-based*—which capture respectively static and dynamic components of user engagement with the topic of

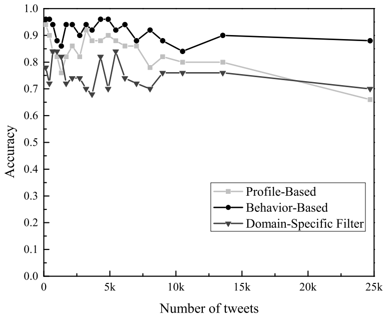


Fig. 5. Dependency of accuracy on the size of available timelines when predicting target users from Twitter dataset.

interest. We also proposed a *domain-specific filter*—a baseline tailored to specific domain used in our case study. We showed that with a slight modification it actually can be extended to other domains.

Experimental results for investment community of Twitter have shown that all three models yield decent performance for a targeted application setting: with *naïve baseline* running in linear time and achieving $F_1 = 0.66$, *behavior-based* obtaining the best results ($F_1 = 0.96$) but being computationally expensive, and the *profile-based* being a trade-off between these two both from time and performance points of view.

We also have discovered that none of these models relies on the size of the timeline of a candidate user, meaning that only fraction of posts can be analyzed even for those very active individuals, this way saving time yet providing declared level of quality.

This framework can be successfully applied to automatic discovery of topical groups on platforms, such as Twitter. Set of selected candidates can be then used for a tailored professional recommendation or selection of an "expert crowd" relevant to external analytical task. For example, for the reported case study content of such experts can be simply treated as a set of recommendations which can be used for devising a sophisticated trading strategy.

Many avenues of research can be considered in the future work. For instance, higher-order models can be explored to model deeper semantic attribution, alternatively variation of traditional topic models can be employed to learn disjoint set of topics characterizing the community. Also *behavior-based* model can be modified the way it actually incorporates explicit temporal analysis. Finally, it is interesting to see how the models perform when applied to different target community.

## Acknowledgment

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

## References

[1] J. Holcomb, J. Gottfried, A. Mitchell, and J. Schillinger, "News use across social media platforms," Pew Research Center, Tech. Rep., November 2013.

[2] L. Willnat and D. H. Weaver, "The american journalist in the digital age: Key findings." School of Journalism, Indiana University, Tech. Rep., 2014.

[3] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[4] W. Liao, S. Shah, and M. Makrehchi, "Winning by following the winners: Mining the behaviour of stock market experts in social media," in *Social Computing, Behavioral-Cultural Modeling and Prediction - 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings*, 2014, pp. 103–110. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-05579-4_13

[5] S. Hill and N. Ready-Campbell, "Expert stock picker: the wisdom of (experts in) crowds," *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 73–102, 2011.

[6] G. Wang, T. Wang, B. Wang, D. Sambasivan, Z. Zhang, H. Zheng, and B. Y. Zhao, "Crowds on wall street: Extracting value from collaborative investing platforms," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, 2015, pp. 17–30. [Online]. Available: http://doi.acm.org/10.1145/2675133.2675144

[7] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, and G. Goldstein, "Identifying and following expert investors in stock microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2011, pp. 1310–1319. [Online]. Available: http://www.aclweb.org/anthology/D11-1121

[8] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[9] I. Becerra-Fernandez, "Facilitating the online search of experts at nasa using expert seeker people-finder." in *PAKM*, ser. CEUR Workshop Proceedings, U. Reimer, Ed., vol. 34. CEUR-WS.org, 2000. [Online]. Available: http://dblp.uni-trier.de/db/conf/pakm/pakm2000.html#Becerra-Fernandez00

[10] D. Yimam-Seid and A. Kobsa, "Expert-finding systems for organizations: Problem and domain analysis and the demoir approach," *Journal of Organizational Computing and Electronic Commerce*, vol. 13, no. 1, pp. 1–24, 2003.

[11] K. Balog, L. Azzopardi, and M. De Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 43–50.

[12] I. Guy, U. Avraham, D. Carmel, S. Ur, M. Jacovi, and I. Ronen, "Mining expertise and interests from social media," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 515–526.

[13] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, "Expertrank: A topic-aware expert finding algorithm for online knowledge communities," *Decision Support Systems*, vol. 54, no. 3, pp. 1442–1451, 2013.

[14] A. Pal, S. Chang, and J. A. Konstan, "Evolution of experts in question answering communities." in *ICWSM*, 2012.

[15] A. Bozzon, M. Brambilla, S. Ceri, M. Silvestri, and G. Vesci, "Choosing the right crowd: expert finding in social networks," in *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 2013, pp. 637–648.

[16] M. Stankovic, M. Rowe, and P. Laublet, "Finding co-solvers on twitter, with a little help from linked data," in *The Semantic Web: Research and Applications*. Springer, 2012, pp. 39–55.

[17] C. Wagner, V. Liao, P. Pirolli, L. Nelson, and M. Strohmaier, "It's not in their tweets: Modeling topical expertise of twitter users," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 91–100.

[18] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman, "Finding news curators in twitter," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 863–870.

[19] P. Bhattacharya, S. Ghosh, J. Kulshrestha, M. Mondal, M. B. Zafar, N. Ganguly, and K. P. Gummadi, "Deep twitter diving: Exploring topical groups in microblogs at scale," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 2014, pp. 197–210.

[20] L. Lü, Y.-C. Zhang, C. H. Yeung, and T. Zhou, "Leaders in social networks, the delicious case," *PloS one*, vol. 6, no. 6, p. e21202, 2011.

[21] Z. Cheng, J. Caverlee, H. Barthwal, and V. Bachani, "Who is the barbecue king of texas?: a geo-spatial approach to finding local experts on twitter," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 335–344.

[22] S. Räbiger and M. Spiliopoulou, "A framework for validating the merit of properties that predict the influence of a twitter user," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2824–2834, 2015.

[23] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011, pp. 45–54.

[24] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi, "Cognos: crowdsourcing search for topic experts in microblogs," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 575–590.