Community Member Retrieval on Social Media using Textual Information

Aaron Jaech, Shobhit Hathi, Mari Ostendorf
University of Washington
{ajaech, shathi, ostendor}uw.edu

Abstract

This paper addresses the problem of community membership detection using only text features in a scenario where a small number of positive labeled examples defines the community. The solution introduces an unsupervised proxy task for learning user embeddings: user re-identification. Experiments with 16 different communities show that the resulting embeddings are more effective for community membership identification than common unsupervised representations.

1 Introduction

Active users of social media often like identifying other users with common interests and values. Or, a user may want to find other users that share characteristics with specific accounts that they follow, e.g. cartoonists or local food trucks. Members of such communities of interest are often identifiable via their social network connections, and shared social connections are clearly important in recommendations. However, shared connections often reflect a subset of a person's interests, and there may be users of interest where any shared connections are distant. In addition, there may be scenarios where there is no explicit social graph, or the full graph is expensive to obtain. In such cases, the language of tweets, blogs, etc. is helpful in identifying users with particular interests.

In this paper, we represent users in terms of the text in their communications and introduce a scenario where a user can define a "community" by providing a small number of example accounts that are used to train a system for retrieving similar users. Note that our use of the term "community" differs from other online contexts, where members explicitly self-identify with a community (e.g. by joining a discussion forum or using a specific hashtag). The community is in the eye of the user issuing the query.

We frame the task of community membership detection as a retrieval problem. A small set of representative accounts selected by the user forms the query, and the system retrieves additional community members from a large index of accounts. The task is loosely related to entity set expansion (Pantel et al., 2009). We make no assumptions about the type of communities that can be handled, and no labeled data is available other than the query. Because the training set (query) is minimal, unsupervised learning is useful for the text representation. We propose the proxy task of person re-identification for learning a user embedding, where the goal is for two embeddings from the same user to be closer to each other than to the embedding of a random user. The hypothesis is that a representation useful for detecting similarities between posts from the same person made at different times will also do well at identifying similarities between people in the same community. This hypothesis stems from observations that people with shared interests often talk about topics related to these interests, and that they tend to have shared jargon and other similarities in language use (Nguyen and Rosé, 2011; Danescu-Niculescu-Mizil et al., 2013; Tran and Ostendorf, 2016).

In this paper, we demonstrate experimentally that the re-identification proxy task is useful with simple models that are suited to the retrieval scenario, and present analyses showing that the approach learns to emphasize words associated with individual interests and polarizing issues.

2 Model

The model for community detection includes: i) a mapping from a user's text (a collection of tweets) to a k-dimensional embedding, and ii) a binary classifier for detecting whether a candidate user belongs to the target community. The novel con-

Proceedings of NAACL-HLT 2018, pages 595–601
New Orleans, Louisiana, June 1 - 6, 2018.

tribution of the work is the proxy re-identification task for learning the user embedding.

User Embedding Model. The mapping from text to an embedding could leverage any document-level representation. We focus on a simple weighted bag-of-words neural model for direct comparison to other popular methods, motivated by the fact that many virtual communities form around shared interests in particular topics. Specifically, let $c_{p,i}$ denote the number of times person p uses word $v_i \in V$, where V is the vocabulary, and $w_{p,i} = \log(c_{p,i} + 1)$ be the log-scaled word count. Then the user embedding is

$$u_p = \frac{w_p^T \mathbf{E}}{||w_p^T \mathbf{E}||} \quad (1)$$

where $w_p = [w_{p,1} \cdots w_{p,|V|}]$ and $\mathbf{E} \in \Re^{|V| \times k}$ is the matrix of word embeddings.

Person Re-identification Learning. The embedding matrix $\mathbf{E}$ is learned using a person re-identification objective that encourages embeddings from the same person to be closer than embeddings from different people. We build on the triplet loss function taken from Schroff et al. (2015) used to train a face recognition system. Specifically:

$$\mathbf{E} = \underset{\mathbf{E}}{\operatorname{argmin}} \sum_{p_1, p_2 \in \mathcal{P}} \text{cost}(p_1, p_2), \quad (2)$$

$$\text{cost}(p_1, p_2) = (1 + \text{d}(u_{p_1^1}, u_{p_1^2}) - \text{d}(u_{p_1^1}, u_{p_2^1}))^+,$$

xwhere $\text{d}(x, y)$ is the cosine distance between x and y. $u_{p_1^1}$ and $u_{p_1^2}$ are embeddings made from distinct subsets of a single person's Tweets, and $u_{p_2^1}$ is an embedding made from a subset of another person's Tweets. In practice, we estimate the loss function randomly sampling triplets (p_1^1, p_1^2, p_2^1) from a large training set.

Classifier. A logistic regression model with L2 regularization is used for the classifier, because it is simple but powerful and our scenario has little training data. Simplicity is important because the classifier should be trainable in real-time after receiving the query. The classifier objective is to discriminate the embeddings from the users in the query from a set of user embeddings from the general collection. For the i-th user, let $y_i \in \{0, 1\}$ be the binary label indicating whether the user belongs to a particular community and u_i be the user embedding. The logistic regression model computes the probability that the user belongs to the community according to:

$$p(y_i = 1|u_i) = \sigma(w^T u_i + b), \quad (3)$$

where $\sigma(x) = 1/(1 - e^{-x})$. During evaluation, the users in the index are ranked according to the maximum log probability ratio

$$\underset{i}{\operatorname{argmax}} \log \frac{p(y_i = 1|u_i)}{p(y_i = 0|u_i)} = \underset{i}{\operatorname{argmax}}\, w^T u_i. \quad (4)$$

Because the classifier is linear, we can quickly retrieve the top matching users from the index using approximate nearest-neighbor search (Kushilevitz et al., 2000). The technique is scalable up to hundreds of millions of users and beyond.

3 Data

All data was collected using the Twitter API.[1] We used 1,035 randomly selected items from the list of trending topics in the USA during the period April-June 2017 to query for users and collected their most recent 2,000 tweets. Example trending topics are #Quantico, RonaldoCristiano, and #MayDay2017. (The full list is available with the data.) Each user had at least one Tweet that mentioned a trending topic but their other Tweets could be on any topic.

We refer to this collection as the "general population," because it was not targeted towards any particular community. In total, we collected around 80,000 such users and used roughly 36,000 for learning user embeddings, 1,000 for learning the community classifiers, and 43,000 for evaluation. The text is mostly in English, but some of it is in Spanish, French, or other languages. A list of the tweet IDs is available.[2]

To support evaluation with the community detection task, we conducted a second collection (contemporaneous with the first) targeting members that we had identified as belonging to one of 16 communities (Table 2). To define a "community," volunteers manually selected a set of users that fit with a theme that they had familiarity with. Thus, the specific 16 communities were determined based on themes of interest to the authors and their friends and colleagues, where we could

[1] http://developer.twitter.com/en/docs/api-reference-index
[2] http://github.com/ajaech/twittercommunities

be reasonably confident about membership decisions. In addition, we tried to avoid themes that might be biased towards well-known celebrities, and we made an effort to have diversity in the characteristics of the communities. The communities were selected to span a range of topics, sizes (6-130 accounts), individuals vs. organizations, and other characteristics. A few of the communities are comprised of organizations rather than individuals such as the high school drama departments and the Pittsburgh food truck communities. (The community names are invented by the authors for purposes of describing the data in this paper; they are not part of the retrieval task.)

The text is lower-cased and some punctuation is removed using regular expressions. Words are formed by splitting on white space. While this strategy will not work for languages that do not delimit words by spaces, these make up a negligible portion of the data. A 174k vocabulary was created by extracting the unique types that were seen in the tweets from the general population, as well as selected bigrams extracted using the open source Gensim library using a point-wise mutual information criteria (Řehůřek and Sojka, 2010). The vocabulary included roughly 49k bigrams, 36k usernames and 17k hashtags. Usernames, hashtags, and URLs are not treated specially and can be part of the vocabulary just like any other word if they occur frequently enough.

4 Experiments

4.1 Experiment Configuration

The experiments involved comparing different methods of learning user embeddings, all with a weighted bag-of-words modeling assumption:

- Weighted word2vec (W2V) using default[3] skip-gram training (Mikolov et al., 2013);
- Latent Dirichlet allocation (LDA) (Blei et al., 2003), using default settings from the Scikit Learn library (Pedregosa et al., 2011);
- Person re-identification with random initialization (RE-ID); and
- Person re-identification with W2V initialization (RE-ID, W2V init).

Both count-weighted W2V and LDA have been used as unsupervised representations in Twitter classification tasks, as noted in Section 5. Default configurations are used because there is insufficient data to have a separate validation set.

For all methods, the same vocabulary, final dimension (128), unit vector normalization strategy, and logistic regression model training were used. The embeddings are trained on the 36k user general data, randomly sampling pairs of users p_1 and p_2 and then sampling 50 tweets at a time without replacement to create $u_{p_1^1}$, $u_{p_1^2}$, and $u_{p_2^1}$. The logistic regression models are trained on the 1K user general training pool, using the 50 most recent tweets for each user. Because there are so few labeled examples for most communities, training and evaluation is done using a leave-one-out strategy with the positive samples but including all of the 1K negative samples. For each of the N classifiers (corresponding to N labeled samples), the test set is the left-out positive example and the 43K general user test pool. Also because of training limitations, there is no tuning of the regularization weight; the default weight of 1.0 is used. Tuning may be useful given a collection of training and testing communities. Performance is averaged over the N classifiers (corresponding to the N labeled samples). Two evaluation criteria are used: a retrieval metric (inverse mean reciprocal rank or 1/MRR) (Voorhees et al., 1999) and a detection metric (area under the curve or AUC).

4.2 Results

Table 1 shows retrieval results averaged across all communities. The RE-ID model outperforms the W2V and LDA baselines for both criteria, with substantial gains in 1/MRR (lower is better). Further, the version of RE-ID initialized with word2vec did better than the one that was initialized randomly even though the randomly initialized version was trained for twice as long.

Strategy	AUC	1/MRR
W2V	93.9	846
LDA	95.0	501
RE-ID (rand. init)	98.0	24
RE-ID (W2V init)	**98.5**	**12**

Table 1: Performance of different model variants.

A breakdown of the best model performance by community is given in Table 2. Sample size does not seem to be a good indicator of performance: the two smallest communities (Cartoonists, Fresno City Council) had the worst and one

[3]The default configuration uses a window of ± 7 words. We also tried using a window of 50 words, which roughly matches the context used in other methods, but community detection performance was significantly worse.

of the best results, respectively. Anecdotally, we observed that the sample of cartoonists were more likely to Tweet about topics outside their main interest (e.g., politics or sports). We hypothesize that the diversity of interests of the members of a community affects the difficulty of the retrieval task, but our test set is too small to confirm this hypothesis.

Community	Size	1/MRR
Cartoonists	8	58.1
Chess Stars	14	5.4
Conan Show Writers	12	4.7
Fashion Commentators	11	8.3
Fresno City Council	6	3.0
Hedge Fund Managers	11	25.7
H.S Drama Departments	18	2.3
Mathematicians	11	32.6
NLP Researchers	50	4.9
Pittsburgh Food Trucks	15	3.3
Police Dogs	16	2.7
Professional Economists	11	3.6
SCOTUS Reporters[4]	16	1.9
The Stranger Reporters[5]	11	8.3
Ultimate Frisbee Players	130	6.7
Ultramarathon Runners	28	14.6

Table 2: W2V+RE-ID results by community

These results may underestimate performance, because there is a chance that some users in the general population test data may actually belong to one or more of our test communities, i.e. there could be mislabeled data. To assess the potential impact, we manually checked the top ten false positives for each community for mislabeled users. We did discover some mislabeled examples for the economist, hedge fund manager, and ultramarathon runner communities. For the most part, the top ranked users from the general population tended to be people from related communities. For example, the top false ultimate frisbee users contained people who wrote about their participation in tournaments for other sports such as soccer.

4.3 Analysis

The finding that the W2V-initialized RE-ID model is significantly better than W2V raises the question: how do the embeddings learned by the re-identification task differ from the ones learned by the word2vec objective? To investigate this, we looked at the 1,000 words in the RE-ID model with embeddings that were farthest (in Euclidean distance) from its word2vec initialization. These top words disproportionately contain Twitter user handles, so some social network structure is captured. Using agglomerative clustering, we found groups of words that centered around frequent words used in particular regions (foreign words, dialects) or cultures (sociolects), associated with hobbies or interests (specific sports, music genres, gaming), or polarizing topics (political parties, controversial issues). At least one of the top tokens was the username of an account later identified as being sponsored by the Russian government to spread propaganda during the United States presidential election, e.g., "ten_gop" in Table 4 of the Appendix.

We also looked at which communities are closest in the embedding space. We represent a community with the average of the member embeddings and use a normalized cosine distance for similarity. The two nearest neighbors are Mathematicians and NLP researchers, which are also close to the next two nearest neighbors, Hedge Fund Managers and Professional Economists.

To interpret what the model as a whole captured, we found the top scoring tweets for each held-out user (creating an embedding for a single tweet) according to the logistic regression model. Representative examples include "recurrent neural_network grammars simplified and analyzed" for NLP Researchers, and "we're looking_forward to seeing you opening_night may 24th love the cast of high_school musical" for High School Drama clubs. Examples for additional communities are included in the appendix. The results provide insight into the community member identification decision.

5 Related Work

One notion of community detection involves discovering different communities within a collection of users (Chen et al., 2009; Di, 2011; Fani et al., 2017). A related task is making recommendations of friends or people to follow (Gupta et al., 2013; Yu et al., 2016). In contrast, our task involves identifying other members of a community, which is specified in terms of a set of example users. These tasks use different learning frameworks (our work uses supervised learning), but the features (social network and/or text cues)

[4]People who write news articles about the Supreme Court of the United States.

[5]The Stranger is a small weekly newspaper.

are relevant across tasks. Our task is perhaps more similar to using social media text to predict author characteristics such as personality (Golbeck et al., 2011), gang membership (Wijeratne et al., 2016), geolocation (Han et al., 2014), political affiliation (Makazhanov et al., 2014), occupational class (Preoţiuc-Pietro et al., 2015), and more. Again, a commonality across tasks is the frequent use of unsupervised representations of textual features.

In representing text, a common assumption is that community language reflects topical interests, so representations aimed at topic modeling have been used, including LDA (Pennacchiotti and Popescu, 2011) and tf-idf weighted word2vec embeddings (Boom et al., 2016; Wijeratne et al., 2016). Yu et al. (2016) compute a user embedding by averaging tweet embeddings. Other work investigates methods for learning embeddings that integrate text and social network (graph or text-based) features (Benton et al., 2016).

The work closest to ours is by Fani et al. (2017), which learns embeddings that are close for like-minded users, where like-minded pairs are identified by a deterministic algorithm that leverages timing of related posts. Our approach requires no additional heuristics for defining user similarity, but instead relies on an objective that maximizes self-similarity and minimizes similarity to other users randomly sampled from a large general pool.

Our person re-identification proxy task makes use of the triplet loss used to learn person embeddings for face recognition (Schroff et al., 2015). In image processing, person re-identification refers to the task of tracking people who have left the field of view of one camera and are later seen by another camera (Bedagkar-Gala and Shah, 2014). It is different from our proxy task and the methods are not the same.

6 Conclusion

In summary, this paper defines a task of community member retrieval based on their tweets, introduces a person re-identification task to allow community definition with a small number of examples, and shows that that the method gives very good results compared to word2vec and LDA baselines. Analyses show that the user embeddings learned efficiently represent user interests. The text embeddings are largely complementary to the social network features used in other studies, so performance gains can be expected from feature combination. While our experiments use a bag-of-words representation, as in most related work, the re-identification training objective proposed here can easily be used with other methods for deriving document embeddings, e.g. (Le and Mikolov, 2014; Kim, 2014).

Acknowledgements

The authors thank the anonymous reviewers for their feedback and helpful suggestions.

References

Apurva Bedagkar-Gala and Shishir Shah. 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–216.

Adrian Benton, Raman Arora, and Mark Dredze. 2016. Learning multiview embeddings of Twitter users. In *Proc. ACL*, volume 2, pages 14–19.

David Blei, Andrew Ng, and Michael Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, 80:150–156.

Jiyang Chen, Osmar R. Zaïane, and Randy Goebel. 2009. Local community identification in social networks. In *International Conference on Advances in Social Network Analysis and Mining*, pages 237–242.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. WWW*.

Ying Di. 2011. Community detection: topological vs. topical. *Journal of Informatics*, 5(4):489–514.

Hossein Fani, Ebrahim Bagheri, and Weichang Du. 2017. Temporally like-minded user community identification through neural embeddings. In *Proc. ACM Conference on Information and Knowledge Management*, pages 577–586.

Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 149–156. IEEE.

Pankaj Gupta, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. Wtf: The who to follow service at Twitter. In *Proc. WWW*, pages 505–514. ACM.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. EMNLP*, pages 1746–1751.

Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. 2000. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474.

Quo Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. ICML*, pages 3104–3112.

Aibek Makazhanov, Davood Rafiei, and Muhammad Waqar. 2014. Predicting political preference of twitter users. *Social Network Analysis and Mining*, 4(1):1–15.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*, pages 3111–3119.

Dong Nguyen and Carolyn P. Rosé. 2011. Language use as a reflection of socialization in online communities. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 76–85. Association for Computational Linguistics.

Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. 2009. Web-scale distributional similarity and entity set expansion. In *Proc. EMNLP*, pages 938–947. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to Twitter user classification. In *ICWSM*.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015. An analysis of the user occupational class through Twitter content. In *Proc. ACL-IJCNLP*, pages 1754–1764.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. CVPR*, pages 815–823.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. In *Proc. EMNLP*.

Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Sanjaya Wijeratne, Lakshika Balasuriya, Derek Doran, and Amit Sheth. 2016. Word embeddings to enhance Twitter gang member profile identification. In *Proc. IJCAI Workshop on Semantic Machine Learning*.

Yang Yu, Xiaojun Wan, and Xinjie Zhu. 2016. User embedding for scholarly microblag recommendation. In *Proc. ACL*, pages 449–453.

Appendix: Supplementary Tables

Community	Selected Tweet
Chess Stars	@chesscom yep karpov well_done twittersphere
Professional Economists	#china real_estate as long as liquidity remains ample this will continue
Fashion Commentators	rihanna's fenty corp creative_director jahleel weaver styles the collection on 3 muses
Fresno City Council	gr8 resource developed by our local @citdfresno on how to export @city-offresno @fresnocountyedc lee_ann eager
High School Drama	we're looking_forward to seeing you opening_night may 24th love the cast of high_school musical
Mathematicians	forms of knowledge of advanced mathematics for teaching (i wrote a thing)
NLP Researchers	recurrent neural_network grammars simplified and analyzed
Police Dogs	when a trained police dog is placed with another handler they complete a re handling course to be licensed normally 2_weeks
SCOTUS Reporters	as supreme_court throws out two gop-drawn congressional_districts as unconstitutional racial gerrymanders
Ultramarathon Runners	we're covering the lake sonoma 50 mile live on saturday tell your friends spread the word and get ready

Table 3: Top tweets for selected communities. Underscore is used to join bigrams.

Interpretation	Top Words
Languages & Dialects	• à, ça, j'ai, quand, c'est, avec, sur, dans_le • é, não, melhor, tem, mesmo, só, mais, hoje, uma, tá, já • es_un, más, jugar, en_el, maduro, jajajaja • bruh, dawg, @iamakademiks, black_women, @chancetherapper, lmaooo, y'all, tryna
Sports	• @mlb, baseball, bullpen, @angels, mets, mlb • arsenal, mate, liverpool, @manutd, mourinho, #mufc • @nhl, hockey, nhl, leafs, @nhlblackhawks, @nhlonnbcsports • xd, @playoverwatch, #ps4share, anime, @keemstar, overwatch, twitch, @nintendoamerica, gaming
Music	• @niallofficial, @harry_styles, @louis_tomlinson, @ashton5sos, @shawnmendes, @ethandolan, @graysondolan, @michael5sos, @danisnotonfire
Political	• @indivisibleteam #resist, #trumpcare, @ezlusztig, @kurteichenwald, @georgetakei, @sarahkendzior, @repadamschiff, @malcolmnance, @lawrence • @mitchellvii, @prisonplanet, @realjameswoods, @jackposobiec, @bfraser747, @cernovich, @ten_gop, #maga
Other	• tories, labour, corbyn, #auspol, tory, mum, nhs, lads scotland

Table 4: Clusters of words that change the most between Word2Vec and the re-identification objective.