



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Social Links from Latent Topics in Microblogs

Citation for published version:

Cohen, S, Puniyani, K, Eisenstein, J & Xing, EP Social Links from Latent Topics in Microblogs.

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Author final version (often known as postprint)

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Social Links from Latent Topics in Microblogs*

Kriti Puniyani and Jacob Eisenstein and Shay Cohen and Eric P. Xing

School of Computer Science

Carnegie Mellon University

{kpuniyan,jacobeis,scohen,epxing}@cs.cmu.edu

1 Introduction

Language use is overlaid on a network of social connections, which exerts an influence on both the topics of discussion and the ways that these topics can be expressed (Halliday, 1978). In the past, efforts to understand this relationship were stymied by a lack of data, but social media offers exciting new opportunities. By combining large linguistic corpora with explicit representations of social network structures, social media provides a new window into the interaction between language and society. Our long term goal is to develop joint sociolinguistic models that explain the social basis of linguistic variation.

In this paper we focus on *microblogs*: internet journals in which each entry is constrained to a few words in length. While this platform receives high-profile attention when used in connection with major news events such as natural disasters or political turmoil, less is known about the themes that characterize microblogging on a day-to-day basis. We perform an exploratory analysis of the content of a well-known microblogging platform (Twitter), using topic models to uncover latent semantic themes (Blei et al., 2003). We then show that these latent topics are predictive of the network structure; without any supervision, they predict which other microblogs a user is likely to follow, and to whom microbloggers will address messages. Indeed, our topical link predictor outperforms a competitive supervised alternative from traditional social network analysis. Finally, we explore the application of supervision to our topical link predictor, using regression to learn weights that emphasize topics of par-

ticular relevance to the social network structure.

2 Data

We acquired data from Twitter’s streaming “Gardenhose” API, which returned roughly 15% of all messages sent over a period of two weeks in January 2010. This comprised 15GB of compressed data; we aimed to extract a representative subset by first sampling 500 people who posted at least sixteen messages over this period, and then “crawled” at most 500 randomly-selected followers of each of these original authors. The resulting data includes 21,306 users, 837,879 messages, and 10,578,934 word tokens.

Text Twitter contains highly non-standard orthography that poses challenges for early-stage text processing.¹ We took a conservative approach to tokenization, splitting only on whitespaces and apostrophes, and eliminating only token-initial and token-final punctuation characters. Two markers are used to indicate special tokens: #, indicating a topic (e.g. #curling); and @, indicating that the message is addressed to another user. Topic tokens were included after stripping the leading #, but address tokens were removed. All terms occurring less than 50 times were removed, yielding a vocabulary of 11,425 terms. Out-of-vocabulary items were classified as either words, URLs, or numbers. To ensure a fair evaluation, we removed “retweets” – when a user reposts verbatim the message of another user – if the original message author is also part of the

*We thank the reviewers for their helpful suggestions and Brendan O’Connor for making the Twitter data available.

¹For example, some tweets use punctuation for tokenization (You look like a retired pornstar!lmao) while others use punctuation inside the token (l0v!n d!s th!ng call3d l!f3).

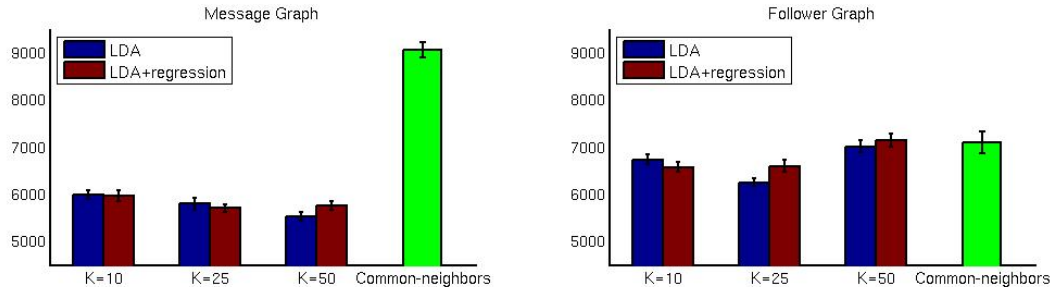


Figure 1: Mean rank of test links (lower is better), reported over 4-fold cross-validation. Common-neighbors is a network-based method that ignores text; the LDA (Latent Dirichlet Allocation) methods are grouped by number of latent topics.

dataset.

Links We experiment with two social graphs extracted from the data: a **follower graph** and a **communication graph**. The follower graph places directed edges between users who have chosen to follow each other’s updates; the message graph places a directed edge between users who have addressed messages to each other (using the @ symbol). Huberman et al. (2009) argue that the communication graph captures direct interactions and is thus a more accurate representation of the true underlying social structure, while the follower graph contains more connections than could possibly be maintained in a realistic social network.

3 Method

We constructed a topic model over twitter messages, identifying the latent themes that characterize the corpus. In standard topic modeling methodology, topics define distributions over vocabulary items, and each document contains a set of latent topic proportions (Blei et al., 2003). However, the average message on Twitter is only sixteen word tokens, which is too sparse for traditional topic modeling; instead, we gathered together all of the messages from a given user into a single document. Thus our model learns the latent topics that characterize *authors*, rather than messages.

Authors with similar topic proportions are likely to share interests or dialect, suggesting potential social connections. Author similarity can be quantified without supervision by taking the dot product of the topic proportions. If labeled data is available (a partially observed network), then regression can be applied to learn weights for each

topic. Chang and Blei (2009) describe such a regression-based predictor, which takes the form $\exp(-\eta^T(\bar{z}_i - \bar{z}_j) \circ (\bar{z}_i - \bar{z}_j) - \nu)$, denoting the predicted strength of connection between authors i and j . Here \bar{z}_i (\bar{z}_j) refers to the expected topic proportions for user i (j), η is a vector of learned regression weights, and ν is an intercept term which is only necessary if a the link prediction function must return a probability. We used the updates from Chang and Blei to learn η in a post hoc fashion, after training the topic model.

4 Results

We constructed topic models using an implementation of variational inference² for Latent Dirichlet Allocation (LDA). The results of the run with the best variational bound on 50 topics can be found at <http://sailing.cs.cmu.edu/socialmedia/naacl10ws/>. While many of the topics focus on content (for example, electronics and sports), others capture distinct languages and even dialect variation. Such dialects are particularly evident in stopwords (*you* versus *u*). Structured topic models that explicitly handle these two orthogonal axes of linguistic variation are an intriguing possibility for future work.

We evaluate our topic-based approach for link prediction on both the **message** and **follower** graphs, comparing against an approach that only considers the network structure. Liben-Nowell and Kleinberg (2003) perform a quantitative comparison of such approaches, finding that the relatively simple technique of counting the number of shared neigh-

²<http://www.cs.princeton.edu/~blei/lda-c>

bors between two nodes is a surprisingly competitive predictor of whether they are linked; we call this approach common-neighbors. We evaluate this method and our own supervised LDA+regression approach by hiding half of the edges in the graph, and predicting them from the other half.

For each author in the dataset, we apply each method to rank all possible links; the evaluation computes the average rank of the true links that were held out (for our data, a random baseline would score 10653 – half the number of authors in the network). As shown in Figure 1, topic-based link prediction outperforms the alternative that considers only the graph structure. Interestingly, post hoc regression on the topic proportions did not consistently improve performance, though joint learning may do better (e.g., Chang and Blei, 2009). The text-based approach is especially strong on the message graph, while the link-based approach is more competitive on the followers graph; a model that captures both features seems a useful direction for future work.

References

- D. Blei, A. Ng, and M. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- J. Chang and D. Blei. 2009. Hierarchical relational models for document networks. *Annals of Applied Statistics*.
- M.A.K. Halliday. 1978. *Language as social semiotic: The social interpretation of language and meaning*. Hodder Arnold.
- Bernardo Huberman, Daniel M. Romero, and Fang Wu. 2009. Social networks that matter: Twitter under the microscope. *First Monday*, 14(1–5), January.
- D. Liben-Nowell and J. Kleinberg. 2003. The link prediction problem for social networks. In *Proc. of CIKM*.

Conversations (out of 10)	Chit-chat Slang Weather Optimism	getting ready go to dinner and movie with some friends!!!!!!!!!! @MsGumdrop lmao das ok now u 1 step smarter den da next person onlything i remember is math n science smh @JPRennquist concerned on the amount of ice that may form over Lake Superior, that would play a bigger role on our weather by March thru May Feeling the hope that a new day brings.	-OOV-, my, to, and, in, with, at, for, the, a, new, going, day, work, so, am, night, today, time, is -OOV-, u, lol, i, me, my, a, 2, im, ur, on, up, lmao, like, in, shit, its, get, got, n the, a, -OOV-, in, it, of, on, is, to, 's, and, cold, for, this, snow, just, here, good, morning, today, weather -OOV-, for, the, of, in, thanks, 's, to, great, at, a, it, out, -URL-, today, new, so, 2, you, see
SPAM (9)	Social media Weight loss	say hi to some new #followfriday frenz: @Kdark74 @MAKEAMILLION1 @janettefuller @The_Scallywags @therealamaru @mslegalhelp Start Losing Weight NOW. Shed holiday pounds safely and effectively. Free Trial Offers. Act today! spon http://tinyurl.com/yd9m8po	you, follow, thanks, ff, for, thank, rt, the, to, back, your, me, are, followfriday, a, great, u, -OOV-, is, hi -URL-, -OOV-, free, weight, to, for, the, on, loss, and, your, get, save, off, a, with, fitness, diet, at, lose, fat
News (6)	Financial Gadgets	Market breadth is very negative; a bearish indication see http://bit.ly/IPlyW #News #Trading #Investing #Finance #SPX #Mkt #Business #Stocks The Switch From iPhone To Android, And Why Your First Impression Is Wrong #gadgets http://bit.ly/4GTPmo	-URL-, for, -OOV-, business, to, in, home, real, the, estate, credit, a, mortgage, on, finance, economy, money, market, and -URL-, -OOV-, the, to, iphone, google, for, and, of, a, on, apple, ces, app, one, 's, in, with, nexus, new
Market (4)	Jobs Gadgets	#jobs Oracle Project Manager 12i in Denver, Colorado http://jobshouts.com/job/10128/ don't miss : Canon PowerShot SD960 IS Digital ELPH Camera (Pink) + 4GB SD Card + Case + NB-4L + Accessory Kithttp://bit.ly/60JFM2	-OOV-, -URL-, job, jobs, for, a, to, me, and, follow, the, in, need, looking, we, php, at, little, rock, i digital, card, 2.0, new, usb, reader, in, one, memory, multi, camera, sdhc, lcd, zoom, mp, -URL-, 2.4, 4x, coffee, pink
Entertain. (3)	Television Film and music	RT @joshmmartin: How about nearly everyone on the Atlanta episode of #americanidol tonight w/ a country accent hasn't been from GA RARE Michael Jackson interview with Soulbeat 1979 http://bit.ly/RBsDk Pls ReTweet & #FollowFriday us ! MJ music 4ever!	ha, -OOV-, la, idol, flu, a, american, boxing, of, the, show, fight, castle, floyd, hee, on, swine, americanidol, you, dude -OOV-, -URL-, the, 's, music, of, new, in, and, a, michael, to, on, movie, for, jackson, with, guitar, by, is
Lifestyle (7)	Prayer Food	Lord someone is in that rubble in Haiti who needs to hang on and believe for a miraculous rescue. Breathe on them!Be their rescue!#tcot#pray Good Food, Good Wine and a Bad Girl: A Slice of Heaven - Chocolate Banana Bread: Chocolate Banana Bread http://bit.ly/5LiSwp	-OOV-, for, god, please, love, and, my, haiti, pray, in, to, the, you, praying, of, her, help, hugs, all, is, jesus -OOV-, -URL-, the, to, and, a, in, for, haiti, food, of, 's, with, our, we, recipe, cooking, on, at, wine

Table 1: Examples from the 50 topics found by the LDA model. There were 8 topics from different languages (French, Dutch, Indonesian, Italian, Portuguese, Spanish, German, mixture of English and Indonesian). Two topics seemed to capture stop-words in different communities, and one topic was a mixture of travel and spam. The remaining 39 topics were manually classified into the six broad categories shown. Sample topics are shown above (a detailed analysis will be found at <http://www.cs.cmu.edu/~kpuniyan/Twitter-topics.html>). In each broad category, the first column roughly summarizes the topic, the second shows a representative tweet from the same topic, and the third column shows the top 20 words in this topic.