Improved Stance Prediction in a User Similarity Feature Space

Improved Stance Prediction in a User Similarity Feature Space

Kareem Darwish[1], Walid Magdy[2], and Tahar Zanouda[1]
[1]Qatar Computing Research Institute, HBKU, Qatar. {kdarwish,tzanouda}@hbku.edu.qa
[2]School of Informatics, The University of Edinburgh, UK. wmagdy@inf.ed.ac.uk

***Abstract*—Predicting the stance of social media users on a topic can be challenging, particularly for users who never express explicit stances. Earlier work has shown that using users' historical or non-relevant tweets can be used to predict stance. We build on prior work by making use of users' interaction elements, such as retweeted accounts and mentioned hashtags, to compute the similarities between users and to classify new users in a user similarity feature space. We show that this approach significantly improves stance prediction on two datasets that differ in terms of language, topic, and cultural background.**

I. Introduction

Stance prediction for online social media users has many interesting applications, such as targeted advertising and polling. However, users may shy away from expressing their stances explicitly or may do so sparingly. Recent work has suggested that using users' tweets and network interactions that are not relevant to the topic at hand can be used in identifying their stance on a particular issue [1], [2], [3]. Such features are effective due to two social phenomena that have been observed in social media, namely homophily and social influence [4], [1]. Homophily is the propensity of individuals to interact with similarly minded individuals, forming smaller social networks. With social influence, attitudes of individuals are affected by the attitudes of others in their social network. In effect, smaller social networks have latent beliefs that manifest themselves in different stances that their members embrace [5].

In this paper, we build on previous work to achieve improved stance prediction using users' non-topically relevant tweets and interactions. However, unlike previous work that uses the text of tweets and users' interactions directly, we apply feature space transformation [6] by employing the similarity between users as classification features in order to infer latent group beliefs, which stem from homophily and social influence. Thus, when predicting the stance of a user at test time, we compute the similarity between the new user and the finite set of users that we had observed during training using a random graph walk with graph reinforcement. Users can be connected to each other using a variety of interaction elements, such as retweeted accounts, shared URLs, or used hashtags. We tested the effectiveness of a variety of interaction elements in transforming the classification task from a text/interaction element feature space into a user similarity feature space. To show the effectiveness of our approach, we experiment with two different datasets that differ in language, topic, users' culture, and construction methodology.

The main contributions of this paper are twofold: (*i*) the transformation of a bag-of-words (words or interaction elements) feature space to a user similarity space (to infer latent group beliefs) using a random graph walk, leading to significant improvements in prediction; and (*ii*) the determination of which interaction elements are effective for stance prediction.

II. Background

Much work has focused on classifying users' political orientation and stance on specific topics [7], [8].Twitter users' political orientation can be deduced based on who they follow [7] or whom they retweet [8], [9]. Many features have been used for Twitter user stance classification such as: tweet text, hashtags, user profile information, and retweeted or mentioned accounts [1], [10], [2]. Rao et al. [11] used socio-linguistic features that include types of utterances (*e.g.,* emoticons and abbreviations) and word n-gram features. They showed that they can distinguish between Republicans and Democrats with more than 80% accuracy. Pennacchiotti and Popescu [2] extended the work of Rao et al. [11] by introducing features based on profile information (screen name, profile description, followers, etc.), tweeting behavior, socio-linguistic features, network interactions, and sentiment. Users tended to form so-called "echo chambers" where they engaged with like-minded users [12], [1], and they also showed persistent beliefs over time and tended to maintain their echo chambers that reveal significant social influence [8], [1], [13]. Duan et al. [14] used so-called "collective classification" techniques to jointly label the interconnected network of users using both their attributes and their relationships. Since there are implicit links between users on Twitter (*e.g.,* they retweet the same accounts or use the same hashtags), collective classification is relevant here. Similarly, Tan et al. [15] showed that using user relationships (follower networks and user mentions) can improve sentiment analysis of Twitter users. SemEval 2016 [16] ran a stance detection shared task. Though we are also performing stance prediction, our proposed technique makes use of users' non-relevant tweets, which are missing from the SemEval 2016 dataset. In the SemEval task, an SVM trained on bag-of-words features served as a strong baseline (we use this as a baseline). Thw participating systems focused on content features and the top system used convolutional neural networks [16].

III. Proposed Method

Our goal is to determine the stance of users using topically non-relevant tweets, thus simulating the situation when a user did not explicitly express a stance. A straightforward method for stance classification is to bundle all tweets for each user into one document, and then to train a stance classifier using word unigrams and bigrams as features. Alternatively, one can use the users' interaction elements as features. Interaction elements in tweets include: links to other users such as mentions, retweets, and replies; links to Web resources (URLs), such as news stories; or engagements with other users through hashtags. The effectiveness of both text and interaction elements was explored in the literature [10], [1], [2], and we use both as baselines. In this paper, we propose using the similarity between users as features, where the similarity between users is computed using interaction elements from their tweets. This is motivated by homophily, where users tend to form subgroups on social networks with common latent beliefs, and the similarity between users can help capture such beliefs. Thus, we compute the similarity between all users, and the feature vector for a user u_i would be the similarity of u_i to all n users in our training set $\{s_{u_i,u_1} .. s_{u_i,u_n}\}$. Since the number of training users would typically be in the tens or in the hundreds at most, the computation of similarities is typically efficient and the resulting feature vector is typically dense. We compute the similarity as the conditional probability $p(u_k|u_i)$ that user u_i would "map" to user u_k. We use a bipartite graph in conjunction with graph reinforcement [17] to estimate the conditional probability. Therefore, given a user u_i, we traverse to interaction element e_j, and then we traverse the graph from e_j to u_k. When traversing from u_i to e_j, we compute $p(e_j|u_i)$ using the maximum likelihood estimate:

$$p(e_j|u_i) = \frac{count_{u_i \ links\_to \ e_j}}{\sum count_{u_i \ links\_to \ \forall e}} \quad (1)$$

Similarly, when traversing from e_j to u_k and computing $p(u_k|e_j)$, we use the maximum likelihood estimate. If u_i and u_k are connected via e_j only, then $p(u_k|u_i)$ would be:

$$p(u_k|u_i) = p(e_j|u_i)p(u_k|e_j) \quad (2)$$

However, since two users could be linked via multiple interaction elements and multiple interactions reinforce the link, we combine the paths as follows:

$$p(u'|u) = 1 - \prod_{\forall u,u' \in users, \forall e \in elements} (1 - p(e|u)p(u'|e)) \quad (3)$$

Equation 2 gives the probability of a single path, and subtracting that probability from 1 gives the probability that the path is incorrect. In equation 3, the product gives the probability that all paths are incorrect, and then subtracts the product from 1 to find the probability that the mapping is correct. Graph reinforcement has the desirable effect of diminishing the contribution of interaction elements that appear with a large number of users, because such elements would be less discriminating between homophilous subgroups. Equation 3 could be modified just to sum the probabilities of all the different paths, but we tested it in side experiments and it led to slightly worse results. Figure 1 illustrates the computation of the mapping probability from user u_1 to users u_2 and u_3. As the example shows, though there is a larger number of links connecting u_1 and u_2, the similarity is lower than that mapping from u_1 to u_3, mainly because of the high number of links between e_b and other users.

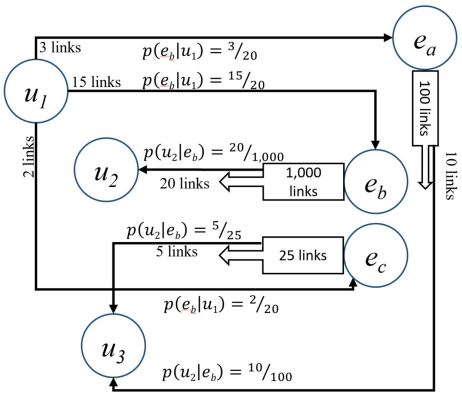

Fig. 1: Example illustrating proposed method for computing the similarity between u_1 and u_2 and u_3 using interaction elements e_a, e_b, and e_c

IV. Experimental Setup

A. Classification Datasets

To test the proposed method, we used two datasets that differ in language, topic, user culture, and construction methodology.

Islands Dataset.: This dataset is part of an on-going project that monitors the attitudes of 21k Arab Twitter users, who are interested in Egyptian politics, from June 2013 to the present. We picked a topic on which many users commented, namely the transfer of ownership of the islands of Tiran and Sanafir from Egypt to Saudi Arabia in April 2016[1]. We filtered the tweets from April 2016 using the Arabic keywords corresponding to: Tiran, Sanafir, Egyptian islands, Saudi islands, two islands, Awwad sold his land (a movie reference), Friday of "the land" (day of protest), "I swear to God they are not our islands", and "King Salman bridge" (a proposed bridge between Egypt and Saudi Arabia via the islands). We allowed phrases to be part of hashtags, with underscores instead of spaces. In all, we found 48,445 matching tweets that were authored by 4,164 users. We submitted all the users along with all their tweets to CrowdFlower to be judged as in favor of the transfer of the islands (POS) or against the transfer (NEG). To ensure quality, we used 50 challenge annotations where annotators have to match the gold annotations we provided to CrowdFlower. All users were judged by three different

[1] http://www.bbc.co.uk/news/world-middle-east-36010965

TABLE I: Islands Dataset Size

	Users	Tweets
POS	687	4,894
NEG	1,777	28,130
Total	2,464	33,024

TABLE II: Islam Dataset Size

	Users	Tweets
POS	2,440	4,974
NEG	972	2,229
Total	3,412	7,203

annotators, and we only retained 2,607 users, who authored 33,207 tweets, where all three annotators agreed on the same judgements. The break down of the different classes is in Table I. These judgements provide the ground-truth stances for the users. For non-topically relevant tweets, we obtained the last 200 tweets that each user authored before April 1, 2016 (before the public discussion began).

Islam Dataset.: This dataset was kindly provided by the authors of [1]. The dataset was collected over the two days after the November 13, 2016 ISIS attacks on Paris using terms such as "#Paris" and "#prayForParis". The tweets were filtered to obtain English tweets originating from the US and mentioning "Islam" or "Muslims". They tagged 979 tweets as expressing positive or negative views towards Muslims. These tweets were retweeted 40,392 times by 35,250 different users who tweeted strictly positive or strictly negative tweets. Since each user (re)tweeted 1.15 tweets on average and many users in the dataset had identical tweets, we decided to randomly pick one user from a set of users with identical tweets. The final set size is shown in Table II. For topically non-relevant tweets, the dataset also includes the last 200 tweets for each user that were authored or retweeted before the attacks.

B. Experimental Conditions

As a baseline for our experiments, we created a document for each user that is composed of the text of its tweets or interaction elements. In all experiments, we randomly pick 100 users for training, and we used the rest for testing. Since the choice of the 100 randomly picked users (*e.g.,* some users have more tweets than others) may affect classification results, for all experiments, we repeated the sample, train, and classify routine 10 times, and we report here the average scores. We used the SVM^{light} implementation of a Support Vector Machine (SVM) binary classifier to classify users as positive or negative [18]. We experimented with the following setups:

Classifying using the Text of the Tweets (TEXT): In this baseline, users' documents were constructed by combining all the text of their 200 topically non-relevant tweets. We used word unigrams and bigrams as features from the combined texts of the tweets. Since the Islands dataset had Arabic tweets, we used a state-of-the-art Arabic stemming and character normalization [19].

Using Interactions Elements as Features (Interaction Elements): For the second baseline, we used user interaction elements from the 200 tweets for each user as features. We used four different interaction elements, namely: (a) retweeted accounts (RT); (b) used hashtags (HASH); (c) mentioned accounts (MEN); and (d) shared URLs (URL). We used interaction element unigrams as features instead of the text features in the TEXT setup.

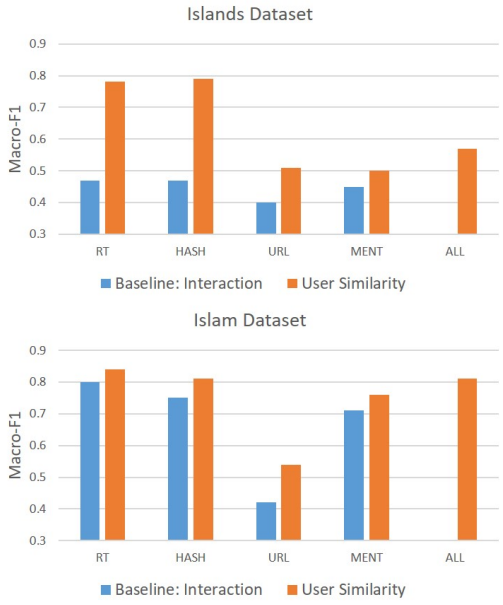

Fig. 2: Comparing baseline and proposed method results for interaction elements for Islands and Islam dataset

Using User Similarity (User Similarity): This represents our proposed method where we used the aforementioned interaction elements individually or collectively (ALL) to compute the similarity between users. Since we used 100 users for training, we computed the similarity between each test user and each of the 100 users in training.

C. Results and Discussion

Table III reports the results when using both baseline conditions and the proposed user similarity features for both datasets. When comparing both baselines, the results show that training on the full text of tweets using word unigrams and bigrams lead to better results compared to using interaction elements for the Islands dataset, but lower results for the Islam dataset. This seems to indicate that the claims in the literature that either the text or the interaction elements are better classification features [1], [2] are not generalizable beyond the test sets they used. The results clearly show that our proposed method, which transforms classification from a text/interaction element feature space into a user similarity feature space, consistently yields improved classification results over both baselines for both datasets. Also as shown in Figure 2, whenever we compare using any interaction element as a feature directly or as a way to compute the similarity, using the interaction element to compute similarity consistently leads to classification improvements with only one exception (mentions

TABLE III: Results of using text and interaction elements baseline and the proposed user similarity.

Islands Dataset											
		Baseline	Baseline: Interaction Elements				User Similarity				
Class	Measure	TEXT	RT	HASH	URL	MENT	RT	HASH	URL	MENT	ALL
POS	P	0.83	0.67	0.66	0.67	0.67	0.83	0.91	0.75	0.74	0.76
	R	0.86	0.90	0.83	1.00	0.92	0.97	0.86	0.97	1.00	0.94
	F1	0.85	0.76	0.74	0.80	0.77	**0.90**	0.88	0.85	0.85	0.84
NEG	P	0.60	0.34	0.32	0.00	0.32	0.90	0.66	0.64	0.88	0.82
	R	0.54	0.11	0.16	0.00	0.07	0.52	0.76	0.10	0.09	0.18
	F1	0.57	0.17	0.21	0.00	0.12	0.66	**0.70**	0.18	0.16	0.30
Macro-F1		0.71	0.47	0.47	0.40	0.45	0.78	**0.79**	0.51	0.50	0.57
Islam Dataset											
		Baseline	Baseline: Interaction Elements				User Similarity				
Class	Measure	TEXT	RT	HASH	URL	MENT	RT	HASH	URL	MENT	ALL
POS	P	0.69	0.85	0.74	0.40	0.53	0.87	0.77	0.79	0.78	0.77
	R	0.62	0.58	0.54	0.00	0.85	0.67	0.68	0.13	0.54	0.68
	F1	0.65	0.69	0.63	0.00	0.65	**0.76**	0.72	0.23	0.64	0.72
NEG	P	0.85	0.85	0.83	0.71	0.92	0.88	0.88	0.74	0.84	0.88
	R	0.89	0.96	0.92	1.00	0.65	0.96	0.92	0.99	0.94	0.92
	F1	0.87	0.90	0.88	0.83	0.76	**0.92**	0.90	0.85	0.88	0.90
Macro-F1		0.76	0.80	0.75	0.42	0.71	**0.84**	0.81	0.54	0.76	0.81

(MENT) for the positive class with F_1 equal to 0.65 when used as feature and 0.64 when used to compute similarity). Often the improvements were quite substantial. For example, using the RT and HASH interaction elements as features led to a macro-F1 of 0.47 and 0.47 respectively compared to 0.78 and 0.79 respectively when using them to compute similarity for the Islands dataset. This may indicate that using user similarity as features tends to capture latent user beliefs that are affected by homophily and user influence. Further, using the RT and HASH interactions elements to compute similarity led to the best results overall, with RT edging HASH for the the Islam dataset and HASH edging RT for the Islands dataset. Using either RT or HASH to compute similarity led to better results than using all of the interaction elements for both datasets. This could be due in part to the poor results of the URL interaction element. Perhaps if we used the names of the sites that the URLs pointed to, this may have led to better results.

V. Conclusion

We have presented an effective method for stance prediction using the similarity between users, which help infer latent group beliefs, as classification features instead of using textual and network interactions directly. The similarity is computed using graph-reinforcement over a random graph walk, where users are connected to each other via interaction elements such as retweeted accounts, mentions, replies, hashtags, and shared URLs. The proposed method yields significant improvement in stance prediction.

References

[1] W. Magdy, K. Darwish, N. Abokhodair, A. Rahimi, and T. Baldwin, "#isisisnotislam or#deportallmuslims?: predicting unspoken views," in *WebSci 2016*, 2016, pp. 95–106.

[2] M. Pennacchiotti and A.-M. Popescu, "Democrats, Republicans and Starbucks afficionados: user classification in Twitter," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 430–438.

[3] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang, "Quantifying political leaning from tweets and retweets." in *ICWSM 2013*, 2013, pp. 640–649.

[4] D. DellaPosta, Y. Shi, and M. Macy, "Why do liberals drink lattes?" *American Journal of Sociology*, vol. 120, no. 5, pp. 1473–1511, 2015.

[5] K. Garimella, "Quantifying and bursting the online filter bubble," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 837–837.

[6] G. Wu and E. Y. Chang, "Adaptive feature-space conformal transformation for imbalanced-data learning," in *ICML*, 2003, pp. 816–823.

[7] P. Barberá, "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data," *Political Analysis*, vol. 23, no. 1, pp. 76–91, 2015.

[8] J. Borge-Holthoefer, W. Magdy, K. Darwish, and I. Weber, "Content and network dynamics behind egyptian political polarization on twitter," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 700–711.

[9] I. Weber, V. R. K. Garimella, and A. Batayneh, "Secular vs. islamist polarization in egypt on twitter," in *ASONAM 2013*, 2013, pp. 290–297.

[10] W. Magdy, K. Darwish, and I. Weber, "#failedrevolutions: Using twitter to study the antecedents of isis support," *First Monday*, vol. 21, no. 2, 2016.

[11] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, 2010, pp. 37–44.

[12] I. Himelboim, S. McCreery, and M. Smith, "Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter," *Journal of Computer-Mediated Communication*, vol. 18, no. 2, pp. 40–60, 2013.

[13] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification." *ICWSM*, vol. 11, no. 1, pp. 281–288, 2011.

[14] Y. Duan, F. Wei, M. Zhou, and H.-Y. Shum, "Graph-based collective classification for tweets," in *CIKM 2012*. ACM, 2012, pp. 2323–2326.

[15] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li, "User-level sentiment analysis incorporating social networks," in *SIGKDD 2011*, 2011, pp. 1397–1405.

[16] S. M. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *SemEval 2016*, vol. 16, 2016.

[17] A. El-Kahky, K. Darwish, A. S. Aldein, M. A. El-Wahab, A. Hefny, and W. Ammar, "Improved transliteration mining using graph reinforcement," in *EMNLP 2011*, 2011, pp. 1384–1393.

[18] T. Joachims, *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer, 2002.

[19] A. Abdelali, N. Durrani, K. Darwish, and H. Mubarak, "Farasa: A fast and furious segmenter for arabic," in *NAACL 2016*, 2016.