Multi-Attribute Topic Feature Construction for Social Media-based Prediction

Alex Morales*, Nupoor Gandhi*, Man-pui Sally Chan*, Sophie Lohmann*,
Travis Sanchez†, Kathleen A. Brady§ Lyle Ungar‡, Dolores Albarracín*, ChengXiang Zhai*
*Dept. of Computer Science and Dept. of Psychology University of Illinois, Urbana-Champaign,
{amorale4, nupoorg2, sallycmp, lohmann2, dalbarra, czhai }@illinois.edu
†Rollins School of Public Health, Emory University
travis.sanchez@emory.edu
§Philadelphia Department of Public Health, AIDS Activities Coordinating Office
‡Computer and Information Science, University of Pennsylvania
ungar@cis.upenn.edu

***Abstract*—The effectiveness of social media-based prediction highly depends on whether we can construct effective content-based features based on social media text data. Features constructed based on topics learned using a topic model are very attractive due to their expressiveness in semantic representation and accommodation of inexact matching of semantically related words. We develop a novel general framework for constructing multi-attribute topic features using multi-views of the text data defined according to metadata attributes and study their effectiveness for a text-based prediction task. Furthermore we propose and study multiple weighting strategies to align text-based features and prediction outcomes. We evaluate the proposed method on a Twitter corpus of over 100 million tweets collected over a seven year period in 2009-2015 to predict human immunodeficiency virus (HIV) new diagnosis and other sexually transmitted infections (STIs) new diagnosis in the United States at the zipcode-level and county-level resolutions. The results show that feature representations based on attributes such as authors, locations, and hashtags are generally more effective than the conventional topic feature representation.**

1. Introduction

The abundance, and ubiquity, of social media data and the live-stream reporting of events make social media data especially valuable for prediction tasks in many application domains.

While there are many applications of social media, using social media for prediction is especially important because it can directly help optimize decision making and can also be combined with other non-text data in a predictive model. As in many cases of text-prediction applications, the accuracy of prediction, based on social media, would highly depend on whether we can construct effective features using the social media data, thus how to construct effective features is an extremely important research question in social media mining. While commonly used features such as bag-of-words representation are often effective, they have clear limitations.

Though promising, a straightforward application of topic modeling to tweets tends to be not very effective. Specifically Twitter, as a source of information, is limited by the message length at 140 characters[1], which restricts the types of content-based features used.

In particular, direct application of a topic model such as LDA [1] to tweets has been shown to produce low-quality topics and thus it is crucial to pool tweets to create coherent documents [2], [3]. However, it remains an open challenge how to pool the tweets and how to construct effective topic-based features to represent tweets in a prediction task, particularly how to determine values of topic features and how to weigh topics for a prediction task.

In this paper we propose a general framework for constructing topics based on social media text data from multiple views that correspond to different ways to pool social media text such as tweets. Those views are defined based on meaningful meta data such as authors, location, and time, each leading to a different, but coherent way of partitioning and pooling text data, and thus enabling generation of coherent topics representing the text data from a different perspective. We call such multi-attribute topic features "discourse features" because each view also can be regarded as providing a "discourse structure" for the text data.

We evaluate the effectiveness of the proposed methods of constructing topic features by using a Twitter corpus of over 100 million tweets collected over a seven year period in 2009-2015 to predict the new diagnosis rates of HIV, gonorrhoea, and chlamydia at different temporal and spatial resolutions in the United States, in particular at the zipcode-level and county-level resolutions. The experimental results

1. As of September 2017, Twitter has extended the length limit to included 280 characters for some select users.

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

show that feature representations based on attributes such as authors, locations, and hashtags are generally more effective than the conventional topic feature representation without considering these multi-view attributes.

2. Related Work

To the best of our knowledge, no previous work has studied how to use meta data to construct multi-view topic features for social media health-based prediction. The closest work to ours is the use of topic modeling for tweets in prediction tasks. In this line, topic modeling has also been employed, with some success, in predicting heart disease mortality at the county-level using Twitter [4] and to analyse the language and personality traits on Facebook [5]. Our work proposes a general framework and multiple new strategies for topic feature construction that are shown to perform better than these ad hoc topic feature construction methods.

Twitter as a useful social media information source has been proven adequate for many health-related tasks such as the prediction of suicide [6], influenza rates [7], asthma-related emergency room visits [8], and HIV rates [6], [9]. Few works have used topic modeling approaches for predicting health-related outcomes [4], [5], [10].

Some studies use specific keywords such as the words "flu", "influenza", and associated symptoms like "high fever" [11] to predict flu and influenza trends. While othes have used dictionary based approaches for HIV prevelance rate prediction [9], [12]. For example, in [9] the authors used two dictionaries related to sexual risk behaviours and attitudes; they classified tweets being drug related or sex related messages, if they contained at least one corresponding risk-related term and finally they used the number of risk-related related tweets as an input feature for a down-stream regression task. We made use of the semantic structure in tweets and built topic models which can be aligned to locations and showed how we can develop features, for predicting HIV and other STIs, which are not limited to a closed-vocabulary approach.

Further, some have proposed different schemes for training the development of new models to improve the topics quality. [2], [3]. Hong and Davison [3] used different aggregation strategies to overcome the short message limitation. They show that the induced topic models are a good feature for classification problems. Alvarez-Melis and Saveski [2] compared of different pooling methods, including at the user, hashtag, and conversations level.

3. Multiview Topic Features

We first present some background information which is needed to understand the proposed new problem of extracting multiview topic feature. For clarity, we often use tweets as examples to illustrate an idea or technique, but the idea and technique are usually general and can be applied to any social media data.

3.1. Background

In text-based prediction, the problem can be described as to predict the value of an interesting variable (e.g., HIV rates of a county) based on the text data associated with the variable (e.g., all the tweets produced by people from a county). Such a prediction task is representative of "big data" applications in general, where the data is leveraged to make a prediction of an interesting variable, which further helps support and optimize decison making.

Topics can be learned from text data in an unsupervised way by using a topic model such as LDA [1]. Specifically, given a set of text documents, topic models, such as LDA, can be used to generate two useful outputs $\mathcal{T} = \{\Theta, \Phi\}$, where Φ is a set of topics, each represented as a word distribution, and Θ is a topic distribution for each document indicating the coverage of each topic in the document.

Normally, when we are concerned with a prediction task based on each document, Φ can be used as the features and Θ can directly provide the weights of all the features for each document. However, such a conventional approach is generally inappropriate for many prediction tasks that are not based on a well-defined single document, which include most prediction applications using social media where we generally have to pool multiple tweets together to form a "document" for prediction. For example, in our prediction task of predicting HIV rates in different counties, we would need to pool all the tweets in a county as a "pseudo document."

3.2. Multi-Attribute Feature Construction

A main challenge in topic weighting is that in many topic modelling applications, there is often a misalignment of the "natural" text document representation and the outcome variables, e.g. a tweet message vs zipcode-quarterly HIV new diagnosis rates. While pooling the data may remedy this issue, those pooled documents may not be topically coherent, or introduce population biases and thus may hurt the prediction performance as we have seen in the experiments.

To address this challenge, we propose a general framework for computing multi-attribute topic representations (called multi-attribute features), which can preserve topically coherent documents and reduce those inherent population biases.

Let d be a document in our document collection $d \in \mathcal{D}$. To construct features for prediction, our text document representation, d, needs to be at the same granularity as the predicting data (outcome variable). For example, if we want to predict the HIV rates at county level, each document d would be all the tweets written by people in a particular county. Such an ad hoc combination of all the tweets makes d incoherent, thus using d as a unit for running a topic modeling would be problematic since there will be noisy co-occurrences that may be picked up by the topic model.

Fortunately, it is often the case that we have more detailed information about these documents available, e.g. authors, which can help us develop better topical features.

Specifically, the document $d = \{a_1, a_2, ..., a_{M_d}\}$, can be viewed as a collection, of size M_d, of some attribute a, i.e. a view of the data; in other words, we say a partitions d. All the tweets that have the same attribute value form a document (indeed a "subdocument") that we would refer to as an *attribute document*, denoted by a_i.

Given a particular attribute a, we can then use all its corresponding attribute documents in the entire data set as units (i.e., as a "document") to run a topic model and generate topics and topic distributions for all the attribute documents, which we denote by $\mathcal{T}_a = \{\Theta^{(a)}, \Phi^{(a)}\}$ with $\Theta^{(a)}$ being the topic distributions and $\Phi^{(a)}$ being the word distributions for all the topics discovered.

Thus for attribute (view) a, we can take all the topics in $\Phi^{(a)}$ each as a feature, and compute the weight of feature k (i.e., topic k) in the feature representation for document d as follows: $\theta_{dk} = P(z = k|d)$ where z is a latent variable indicating the topic in document d. Since an attribute forms a partition we can marginalize over the attribute documents,

$$P(z = k|d) = \sum_{a_i \in d} P(z = k|a_i, d)P(a_i|d)$$

where $P(z = k|a_i, d)$ is the topic weight for a partition of d by attribute value a_i that we can directly obtain from $\Theta^{(a)}$. The last term $P(a_i|d)$ signifies the weight of attribute a_i in d, and we will discuss how to set this weight below.

3.3. Discourse Feature Weighting

First, we can consider *Balanced topic weight* (BTW), which is defined as, $P(a_i|d) = \frac{1}{M_d}$ In such a weighting method, we view every distinct value of attribute a as equally important, thus avoiding any bias we might have due to non-uniform amounts of text data contributed by different attribute values (e.g., some authors may have written far more tweets than others, but should not dominate in the representation).

Sometimes, the amount of tweets belonging to each attribute value does matter (e.g., if there are more tweets belonging to one hashtag than another, we might want to retain this difference). To accommodate such a need, we further introduce *proportional topic weight* (PTW), which is defined as, $P(a_i|d) = \frac{|a_i|}{\sum_{j=1}^{M_d} |a_j|}$ In PTW, we see that attributes with more text data would be weighted higher.

Finally, we may also have the *unweighted probability distributions* (UPD), defined as, $P(z = k|d) \propto \sum_{a_i \in d} P(z = k|a_i)$ This weighting scheme encodes directly corpus-wide statistics, since there is no re-weighting of the attribute topic-document distributions.

Note that depending on the attribute a, it is possible that proportional, balance, and unweighted topic weights could be equal. The different pooling schemes for text document representations of the tweet messages we explored are as follows, a *single tweet message*, *pool tweets in a location*, *pool tweets by a single user*, and *pool tweets by hashtags*. Note that for the location we used both zipcodes and counties.

It is worth pointing out that the proposed multi-attribute topic features can also be constructed when there is no naturally available attribute. While our methods are applied in the context of topic modeling, the approaches can also be used to amalgamate any numerical feature which is used for prediction, such as term-frequency counts.

4. Data Sets

CDC STIs corpus. The county-level HIV, chlamydia (CHLA), and gonorrhea (GONO) new diagnosis data are obtained from the Centers for Disease Control and Prevention (CDC) and AIDSVu [2]. In the odd columns of Figure 1 we show the new diagnosis rates via each state in 2014, note the blank regions in the figure represents the suppressed data.[3]

Philadelphia HIV New Diagnosis Dataset. We obtained zipcode-level HIV diagnosis rates per 100,000 from Philadelphia, Pennsylvania which the HIV data included only people aged 13 and older. Data from regions with less than 5 new HIV diagnoses per year or less than 100 inhabitants are routinely suppressed by the CDC, and this suppression criteria were also applicable for the present analysis.

4.1. Twitter Data

Our Twitter corpus ranges from June 2009 to March 2010, November 2011 to December 2015. In total there were more than 3.4 billion tweets, including re-tweets. However in order to use this dataset at the spatial granularity of the STI new diagnosis rates we geotagged our Twitter corpus to zipcodes, and counties, in the United States. The user geotagging problem has been well studied [13]. In this study we developed a heuristic to quickly, and accurately, geotag tweets at the county and zipcode resolutions.

Geo-location. Tweets may contain geo-coordinates, e.g GPS, which we refer as coordinate data for short, and/or a "location" in the meta-data, we refer to location only data. We handle these two geotagging tasks separately, first we describe coordinate mapping and then location mapping: the mappings of those tweets without the coordinate information. This approach is adapted from [14], in which select cities are mapped to counties if they contain at least 95% of the population of all the cities with the same name. A complete description of the geotagging method and performance can be found in [10].

5. Experiment Procedure and Results

The main purpose of our experiments was to examine two basic questions: 1) Is the proposed multiview attribute

2. http://aidsvu.org/

3. Data are estimated for persons aged 13 and older living with an HIV infection diagnosis as of December 31st, of each respective year. Denominators used to calculate rates for county populations were obtained from the U.S. Census Bureaus census estimates for each respective year.

topic features more effective than the regular topic features (which are usually generated using one view). 2) Which of the proposed weighting functions performs the best? These questions can be answered by comparing multiple runs with appropriate parameter configuration. As the baseline single view can be regarded as a special case of the proposed multi-view framework, the baseline method can be easily simulated by restricting to one view (e.g., pooling all tweets in a county), i.e. the natural document representation.

5.1. Data Pre-Processing

We selected three states from our CDC STI corpus which have higher level of STI new diagnosis rates compared to the rest of the country, i.e., these were California, Florida, and New York. We also included Pennsylvania for comparison with our Philadelphia analysis. We log-transformed and standardized these rates. Due to the quarterly nature of the Philadelphia HIV new diagnosis dataset, we included this time resolution for each attribute document representation of the Twitter data. We used a location based representation, such as zipcodes, then construct the four attribute documents, i.e. tweet messages are grouped by quarter belonging to the same zipcode and corresponding to a HIV diagnosis rates. For all of our experiments we used LDA for topic modeling feature construction, normalized our discourse features and used an estimator, fitted on randomized decision trees (extra-tres) [15] for our regression problem. To ensure there were no outliers in the Twitter dataset, we included the attribute documents, whose lengths (e.g. number of tweets) were within three standard deviations of the mean, and we used all of the available new diagnosis testing data in order to compare the document representations. In particular, we only noticed the presence of outliers when considering the authors, which follows a Zipfian distribution, i.e. a right skewed long tailed distribution and only excluded six authors which we manually verified were attributed to spam accounts.

5.2. Result

5.2.1. CDC STIs Diagnosis County-level Prediction. We use the datasets prior to 2013 as training and considered the STI diagnosis for 2014 as the testing dataset. While the per-year STI diagnosis rates are only reported once a year, the tweets have a creation time-stamp which allows us to pool messages by time, in particular we selected at a quarterly temporal resolution with all our attributes

We propose a simple baseline, where all the messages pooled in a county for the entire year of 2014 is a document from which we constructed topics, which simulates a natural pooling strategy. This is a special case of our model where there is only a single attribute encompassing the entire document. We compared the topic features constructed using attributes with this baseline to see if multiview topic features are indeed beneficial.

The training and testing sizes as well as the prediction mean-squared errors (MSE) are shown in Table 1. We applied a two sample t-test comparing the attribute document and weighting scheme result with the baseline and noted results with significant improvement over the baseline or significant decrease in performance compared to the baseline.

We observe that UPD obtains the minimum MSE, which is not too surprising since the diagnosis rates tend to be concentrated in the metropolitan areas as shown in Figure 1, and UPD was constructed to favor populous locations. We also see that the BTW under the author attribute always improves over the baseline. Partitioning by time helps when the training dataset is small, even though HIV new diagnosis for the states is the most sparse of all STI new diagnosis, we can still achieve good performance with the Quartely attribute document. Gonorrhea new diagnosis rates are the most difficult to predict, especially in California which only by using authors and the BTW scheme can we out perform the baseline. Overall using the attributes message and authors yield the best results in particular authors in Florida and California, which have a non-uniform STI-rates distribution and messages were best for Pennsylvania and New York which tend to be more mostly uniform, with few peaks.

5.2.2. Philadelphia Zipcode-level Prediction. Using the available data prior to 2015 (2009-2014) as our training dataset and for the testing data we choose the most recent HIV new diagnosis data in 2015. We tuned our parameters on a development set, which included the Philadelphia zipcode 2014 HIV new diagnosis data for evaluation and the data prior as the training dataset. The training data contained 352 entries, of which 156 were non-missing, and the test data contained 74 entries of which 44 were non-missing.

Attribute Document	Weighting Schemes	Errors	
		mean SE	median SE
Zipcodes	PTW/ BTW/ UPD	18.32	6.01
Author	PTW	15.87	10.24
	BTW	14.93	9.80
	UPD	18.07	10.40
Hashtag	PTW	19.68	14.75
	BTW	19.86	14.75
	UPD	22.77	16.00
Message	PTW	16.64	9.42
	BTW	16.63	9.67
	UPD	17.81	8.21

TABLE 2: Overall HIV new diagnosis prediction results by weighting scheme

We used both the mean squared error (MSE) and median squared error as our error metrics for the Philadelphia prediction. We compare our weighting scheme in Table 2, by predicting the HIV new diagnosis rates directly for each zipcode. A clear pattern from these results is that the UPD performed the worst in almost all cases. The UPD scheme distributes the topic weights to the populous locations and thus relying on having enough tweet messages to represent this distribution.

While both PTW and BTW outperform UPD, both schemes are similar in performance. But when considering authors as attribute documents, BTW has an overall better MSE score than the other schemes. Such results indicate that partitioning by authors works consistently well, since it

Attribute Document	Weighting Schemes	STIs											
		HIV New Diagnosis				Gonorrhea				Chlamydia			
		Florida	California	Pennsylvania	New York	Florida	California	Pennsylvania	New York	Florida	California	Pennsylvania	New York
Baseline	—	0.371	0.243	0.313	0.183	0.461	0.300	1.023	1.033	0.144	0.141	0.259	0.150
Quarterly	PTW	0.311	0.203	0.339	0.221	0.511	0.432	1.015	1.150	0.178	0.119	0.232	0.168
	BTW	0.399	0.238	0.262	0.277	0.536	0.381	0.940	1.222	0.190	0.146	0.205	0.182
	UPD	0.381	0.202	0.366	0.236	0.592	0.318	0.941	1.333	0.203**	0.100	0.199	0.183
Authors	PTW	0.325	0.228	0.258	0.200	0.413	0.443**	0.822	0.692*	0.146	0.110	0.165*	0.100*
	BTW	0.300	0.176	0.248	0.126	0.377	0.283	0.882	0.733	0.143	0.104	0.196	0.088*
	UPD	0.207*	0.137*	0.191	0.116*	0.354	0.325	0.736*	0.610*	0.107	0.103	0.155*	0.086*
Messages	PTW	0.296	0.172	0.292	0.154	0.379	0.414	0.794	0.620*	0.129	0.093	0.134*	0.082*
	BTW	0.274	0.174	0.307	0.152	0.364	0.392	0.819	0.624*	0.129	0.100	0.167*	0.077*
	UPD	0.228*	0.147*	0.180	0.114*	0.408	0.471**	0.638*	0.584*	0.140	0.073*	0.163*	0.095*
Hashtags	PTW	0.319	0.150*	0.245	0.170	0.403	0.480**	0.841	0.690*	0.124	0.087*	0.160*	0.106
	BTW	0.321	0.183	0.305	0.135	0.377	0.561**	0.854	0.581*	0.143	0.107	0.167*	0.101*
	UPD	0.248	0.145*	0.200	0.146	0.365	0.515**	0.838	0.658*	0.137	0.104	0.155*	0.087*
Train Size		228	168	135	150	304	234	256	260	308	267	320	297
Test Size		44	33	27	30	64	57	63	60	64	58	67	61

TABLE 1: Prediction MSEs for 3 county STIs new diagnosis, for four states with our proposed feature construction methods. A $*$ implies significant improvement with $\alpha = 0.1$, and $**$ is significant decrease with $\alpha = 0.1$ over the baseline.

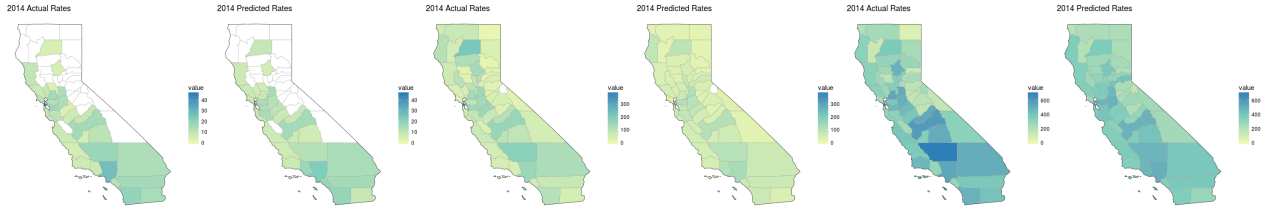

Figure 1: We only show California due to space limitations. Left most two columns: HIV New diagnosis, Middle: Gonorrhea, Right: Chlamydia, predictions for 2014 incident rates, via Authors and UDP scheme.

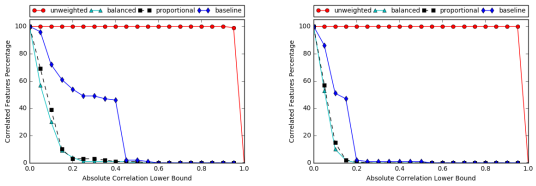

Figure 2: Feature-Message correlations for FL, and CA respectively, using the Author attribute.

avoids the bias from dominance by authors who wrote many more tweets than others (i.e., less biased due to variable data size).

5.2.3. Topic Features Population Bias. We have previously alluded to the population-bias as the effect of depending on message count statistics to produce useful features. We measured this population bias for the CDC STIs county-level prediction by computing the Pearson correlation coefficient with respect to each discourse topic feature and the county tweet message counts. We plot the absolute correlation lower bound and the percentage of features which have a correlation coefficient, whose absolute value is greater than the lower bound for the author attribute and for the GONO testing dataset in Figure 2, e.g. at lower bound of 0; all of the topic features are shown and no feature has above a correlation coefficient of 1. Although not shown the other attributes follow a similar pattern.

We observe that our UDP indeed creates features which are population biased, having a strong message count correlation with more than 90 of all the features. It is also interesting to note that the baseline has about 40 features with a weak correlation (0.2-0.4) for all states except California. Both BTW and PTW do not show this type of association and tend to champaign at 0 before the baseline. We find a similar association with the zipcode features as well. Thus depending on the prediction problem constructing predictive features, UPD could be useful, however if we are interested in making a more robust feature, invariant to the number of messages in some attribute, then it may be better to use the BTW scheme while sacrificing some prediction accuracy.

Discourse Feature Attribute Comparison. While the author discourse features tend to work better with smaller training sample sizes, using messages discourse features in general will work well. It is some what surprising that hashtags do not perform quite on par as authors since, when pooling by hashtags we can expect to create coherent documents. One explanation could stem from the fact that there are many infrequent, as well as very popular hashtags thus causing some disparity in the document sizes. Another factor could be that hashtags are more susceptible to the language shift, since there could be many new events specific to 2015. Thus to measure the topic cohesion we compute the log perplexity of the attributes.

	Quarterly	Message	Author	Hashtags
Log Perplexity	231.89	20.74,	25.07	22.50

TABLE 3: Log Perplexity for different document attributes.

Table 3 presents the log perplexity, so called per-word likelihood bound, for all four different document attributes, Counties, Messages, Hashtags and Authors. To compute the perplexity, we used a withheld development dataset consisting of location only mapped tweets, meaning could not map to zipcodes, instead we used county based map-

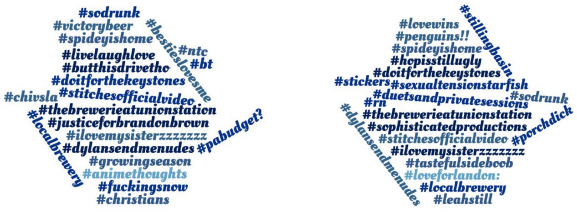

Figure 3: The highest weighted, top 2 topics for Philadelphia zipcodes, with the top-20 highest weighted hashtags, using the UDP scheme.

pings for each quarter in 2015. to construct the document attributes. The perplexity for the quarterly attribute is much worse that the rest, which could be expected since pooling based on time may not necessarily create the most coherent documents. While Hashtags and Messages fit better the development data, it doesn't mean that this is able to translate to the predictive accuracy.

Discourse Feature Analysis. As a qualitative study we show the hashtags discourse topic features in word-clouds, see Figure 3 in order to better observe the topic clusters. We used the topic predictor weights, obtained from our learning algorithm, and selected the top-2 weighted topics, based on the Philadelphia dataset, we then ranked the hashtags themselves based on their weights for these topics and selected the top 20 STI-related hashtags in Figure 3. To identify the STI-related hashtags we used a manually curated STI-related terms to filter hashtags which contain these terms. The hashtags in Figure 3 are all within the top 10% highest ranked hashtags. We find that many indeed are related to sexual themes, e.g. #casualsexweek, but further study is needed to understand in what context and if it is indicative of risky behavior.

6. Conclusion

In this paper, we address a fundamental problem in all those prediction applications, i.e., how to construct effective topic features and proposed a novel framework for constructing multi-view topic features by leveraging a topic model as a building block. The multi-view topic features are constructed based on the multiple attributes of social media data that are naturally available and can be regarded as discourse features. We propose and study three different weighting scheme methods for our discourse features, i.e., unweighted, balanced and proportional, each make different underling assumptions about how the data is distributed and act as regularization methods.

We evaluated the proposed methods using an application on the public health domain – prediction of STIs using tweets, and showed pooling by attributes, such as authors, outperformed the baseline in prediction. The results show that attribute-based multi-view topic features are consistently more effective than the baseline single-view features.

Although the framework is proposed for social media-based prediction, it is general in that the attributes can be defined based on any meta-data available in text-based prediction applications. As the proposed framework is general, another very interesting direction for future work is to explore the application of the general framework in other social media domains.

Acknowledgments

This work was supported by the National Institutes of Health, Grant 1 R56 AI114501-01A1.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[2] D. Alvarez-Melis and M. Saveski, "Topic modeling in twitter: Aggregating tweets by conversations." in *ICWSM*, 2016, pp. 519–522.

[3] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*. ACM, 2010, pp. 80–88.

[4] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap *et al.*, "Psychological language on twitter predicts county-level heart disease mortality," *Psychological science*, vol. 26, no. 2, pp. 159–169, 2015.

[5] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. Seligman *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PloS one*, vol. 8, no. 9, p. e73791, 2013.

[6] J. Jashinsky, S. H. Burton, C. L. Hanson, J. West, C. Giraud-Carrier, M. D. Barnes, and T. Argyle, "Tracking suicide risk factors through twitter in the us," *Crisis*, 2015.

[7] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic," *PloS one*, vol. 6, no. 5, p. e19467, 2011.

[8] S. Ram, W. Zhang, M. Williams, and Y. Pengetnze, "Predicting asthma-related emergency department visits using big data," 2015.

[9] S. D. Young, C. Rivers, and B. Lewis, "Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes," *Preventive medicine*, vol. 63, pp. 112–115, 2014.

[10] M. Chan, S. Lohmann, A. Morales, C. Zhai, L. Ungar, D. Holtgrave, and D. Albarracín, "An online risk index for the cross-sectional prediction of new hiv chlamydia, and gonorrhea diagnoses across us counties and across years." *AIDS and behavior*, 2018.

[11] J. C. Santos and S. Matos, "Analysing twitter and web queries for flu trend prediction," *Theoretical Biology and Medical Modelling*, vol. 11, no. Suppl 1, p. S6, 2014.

[12] M. E. Ireland, Q. Chen, H. A. Schwartz, L. H. Ungar, and D. Albarracin, "Action tweets linked to reduced county-level hiv prevalence in the united states: Online messages and structural determinants," *AIDS and Behavior*, vol. 20, no. 6, pp. 1256–1264, 2016.

[13] B. Han, A. Hugo, A. Rahimi, L. Derczynski, and T. Baldwin, "Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text," *WNUT 2016*, p. 213, 2016.

[14] M. E. Ireland, H. A. Schwartz, Q. Chen, L. H. Ungar, and D. Albarracín, "Future-oriented tweets predict lower county-level hiv prevalence in the united states." *Health Psychology*, vol. 34, no. S, p. 1252, 2015.

[15] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.