A Probabilistic Geographical Aspect-Opinion Model for Geo-tagged Microblogs

Aman Ahuja
Virginia Tech
Blacksburg, VA, USA
aahuja@vt.edu

Wei Wei
Carnegie Mellon University
Pittsburgh, PA, USA
weiwei@cs.cmu.edu

Wei Lu
Singapore University
of Technology & Design
Singapore
luwei@sutd.edu.sg

Kathleen M. Carley
Carnegie Mellon University
Pittsburgh, PA, USA
kathleen.carley@cs.cmu.edu

Chandan K. Reddy
Virginia Tech
Arlington, VA, USA
reddy@cs.vt.edu

***Abstract*—Due to the rapid increase in the number of users owning location-based devices, there is a considerable amount of geo-tagged data available on social media websites, such as Twitter and Facebook. This geo-tagged data can be useful in a variety of ways to extract location-specific information, as well as to comprehend the variation of information across different geographical regions. A lot of techniques have been proposed for extracting location-based information from social media, but none of these techniques aim to utilize an important characteristic of this data, which is the presence of aspects and their opinions, expressed by the users on these platforms. In this paper, we propose Geographic Aspect Opinion model (GASPOP), a probabilistic model that jointly discovers the variation of aspect and opinion, that correspond to different topics across various geographical regions from geo-tagged social media data. It incorporates the syntactic features of text in the generative process to differentiate aspect and opinion words from general background words. The user-based modeling of topics, also enables it to determine the interest distribution of various users. Furthermore, our model can be used to predict the location of different tweets based on their text. We evaluated our model on Twitter data, and our experimental results show that GASPOP can jointly discover latent aspect and opinion words for different topics across latent geographical regions. Moreover, a quantitative analysis of GASPOP using widely used evaluation metrics shows that it outperforms the state-of-the-art methods.**

***Keywords*-Microblogs, probabilistic models, aspect mining, opinion mining, topic modeling.**

I. Introduction

Micro-blogging services, such as Twitter, have become an indispensable mode of communication in recent years. The increase in the number of people using these platforms to express their opinions regarding products and services, as well as their views regarding the political and regional issues concerning them, makes these websites a rich source of information to determine public opinion. Aspect-opinion mining, which aims to discover opinions about different aspects related to an entity, such as a product or service, has become an interesting line of research in the field of text mining. Opinion mining [1] from social-media platforms is also being extensively used by companies to ascertain the public opinion regarding their products and services. Similarly, it can be useful in an election to determine the public support towards different candidates.

With the recent increase in the number of users accessing these services from mobile devices, websites like Twitter and Facebook now allow their users to share their location along with social media posts, either by allowing them to explicitly specify the location or by embedding their geographical coordinates in the posts. This has led to the availability of an enormous amount of geo-tagged data on these platforms, which can be vital to extract meaningful location-specific information, such as event-detection.

Need for geo-spatial aspect-opinion mining: An interesting research problem is: how to utilize the geolocation information embedded in social media posts, along with their text, to discover how various aspects associated with a topic (and their corresponding opinions), vary across geographical regions. This can be useful in a variety of applications, such as to determine:

- *Public interest in a domain:* the interest of people in any domain, such as sports or politics, varies across geographical regions. People living in United States might be interested in sports such as baseball and basketball, whereas in European countries, people might be more interested in soccer.
- *Acceptance of a new law passed by the government:* the new regime might be beneficial for people belonging to few states, and people from these states will have positive opinion about that law. However, it might be detrimental for people from other states, which will be reflected by negative opinion towards this topic in these states.

Overview of the proposed approach: In this paper, we propose **G**eographic **ASP**ect **OP**inion model (GASPOP), to address the problem of discovering latent topics, their aspects and aspect-specific opinions, and their variation across different geographical regions, from geo-tagged microblogs. For each topic, GASPOP model aims to discover its general words, and the corresponding aspect and opinion words in different geographical regions. It utilizes geolocation (*latitude, longitude*) data in these social media posts to discover different latent geographic regions based on the text information in these posts. Then, it aims to determine how different topics, i.e., their aspects and opinions, vary across these geographical regions. To classify words as aspect or opinion, it incorporates syntactic features like part-of-speech (POS) tags, which is the grammatical category (noun, verb, etc.) of the words, in the generative process. The model can also be used to predict the

location of a tweet, based on its text, with considerably higher accuracy than baseline approaches.

II. Related Work

A. Topic Modeling

Topic modeling techniques, which aim to discover latent topics from text, have been widely studied in the field of text mining. Probabilistic Latent Semantic Indexing (pLSI) [2] was one of the earliest techniques to discover topics from text documents, by representing each document as a mixture of topics. Inspired by the success of Latent Dirichlet Allocation (LDA) [3] in mining latent representations of text, a lot of topic modeling techniques have been proposed to discover latent topics in social media data. An important characteristic of social media text is *short-length*, which makes the traditional LDA-based topic models ineffective while mining topics in microblogs. [4], [5] take into account this property, and assign a single topic to all the words in a tweet. These models also use a user-based modeling, which enables them to discover the interests of social media users. Few other works also developed techniques that incorporate the user-based modeling of Twitter data [6], [7]. Recent work in the field of topic modeling focuses on multi-dimensional topic models, where a document is considered to be a combination of words generating from multiple language models [4], [8], [9]. Such a distribution ensures that the commonly used words are separated from the topic words, to give more meaningful topics.

B. Aspect Opinion Mining

Some of the earliest techniques proposed for opinion mining used customer reviews for products [10], [11]. Most of the recent work in the field of aspect-opinion mining uses LDA-based models to discover latent aspects and opinions from review datasets. The ME-LDA model [12] aims to model both aspects and their corresponding opinion words, using a feature vector composed of POS tags. These techniques do not work well in case of microblog text, where the text is less-structured as compared to review text. Opinion mining from Twitter was first studied in [13], using a Naive Bayes classifier to extract opinion from tweets. Later, [14], [15] discussed the use of Twitter data for opinion mining in political elections. These techniques use sentiment classification of individual tweets, and cannot explicitly discover different topics and their opinions from a large corpora of text. The ASEM model [16] is a recently proposed technique that models aspect and opinion words corresponding to different events from Twitter corpora. Similar to ME-LDA, ASEM also uses the POS tags of words to distinguish between aspect and opinion words. However, it does not have a geospatial modeling of events, that could be used to understand the geographical attributes of events.

C. Geographical Modeling of Text

A pLSI-based spatio-temporal model to discover latent topics (or *themes*) as well as location-specific topics for a given time period is proposed in [17]. However, the model does not assume a prior distribution for topics. Moreover, it assumes that the data is already partitioned into geographical regions and time intervals. [18] also assumes that the data belongs to predefined regions and does not incorporate its exact geographical coordinates in the generative process. The GeoFolk model [19] was the earliest work that explicitly incorporated the actual spatial coordinates in the generative process using a Normal distribution on the coordinates. The Geographic Topic Model [20] introduces the concept of latent regions, and aims to discover the geographical variation of different topics across various regions. However, it does not assume a dependency between the latent topics and regions. [21] takes into account this dependency between different topics and regions. It also incorporates user-dependent information in the generative process. Finally, location-based modeling of Twitter data has also been discussed in [22].

III. The Proposed Model

In this section, we introduce GASPOP, a probabilistic model which aims to discover the aspects and their opinions, associated with topics across different geographical regions, from geo-tagged microblogs, like tweets. The model is based on the following observations:

- Due to the short-length of microblogs, each document is associated with only one topic, which depends on the interest of the user (author).
- A topic might be portrayed differently by people from different locations. Moreover, a topic in one location might be irrelevant to people from another location.
- For a topic relevant to multiple regions, people from different regions might be interested in different aspects of the topic. For example, in sports domain, people in USA might be interested in baseball, while people in India might be interested in Cricket.
- People from different locations might have different opinions towards a topic. For example, in an election, a candidate might have support in one region, but not the other.

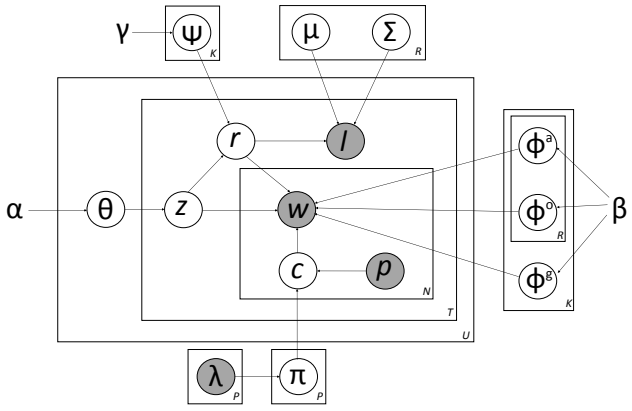

Fig. 1: Plate notation of GASPOP model.

A. Model Definition

GASPOP is a generative model that can discover latent topics, their aspect, and aspect-specific opinion words, as well

as their distribution across different geographical regions from geo-tagged social media data. The plate notation of this model is illustrated in Figure 1.

- The model assumes that there are *K* topics, *R* geographical regions, and *U* users, where the values of *K* and *R* are fixed.
- It models each user *u* as a mixture of topics. For each tweet *t* by the user *u*, a topic is drawn based on the user's interest distribution θ_u. Each geographical region *r* in the model is represented by a geographical center μ_r and covariance Σ_r.
- Each topic *k* in the model is represented by three language models. The general model ϕ^g represents the set of general words that are common across all geographical regions for a topic *k*. To model region-specific aspect and opinions, the model uses an aspect model ϕ^a and an opinion model ϕ^o.
- To distinguish between general, aspect and opinion words, the model uses a category variable $c_{utn} \in \{general, aspect, opinion\}$ for each word in the tweet (*u, t*), whose value depends on the POS tag p_{utn} of the word.

Using Part-of-Speech tags for word category identification: In order to classify a word as aspect, opinion, or general, GASPOP model uses the part-of-speech tags of words in the generative process. This is done to take into account the dependencies between the POS tag and the category of the word; for example, nouns are more likely to be aspect words, whereas adjectives are more likely to be opinion words. This prior knowledge can be incorporated into the value of the hyperparameter λ. In contrast to ME-LDA, where the model also uses the POS tags of the adjacent terms to determine the category of the word, we only use the POS tag *p* of the current word, and a full Bayesian distribution on *c*. This is because tweets usually contain a lot of abbreviations and mis-spelled words owing to their short length.

Algorithm 1: Generative Process of GASPOP

for *each POS tag p* **do**
  Draw category distribution $\pi_p \sim Dirichlet(\lambda_p)$
end
for *each topic z* **do**
  Draw region distribution $\psi_z \sim Dirichlet(\gamma)$
  Draw general words distribution $\phi^g_z \sim Dirichlet(\beta)$
  for *each region g* **do**
    Draw aspect words distribution $\phi^a_{z,g} \sim Dirichlet(\beta)$
    Draw opinion words distribution $\phi^o_{z,g} \sim Dirichlet(\beta)$
  end
end
for *each user u* **do**
  Draw topic distribution $\theta_u \sim Dirichlet(\alpha)$
  for *each tweet t by user u* **do**
    Draw topic $z_{ut} \sim Multinomial(\theta_u)$
    Draw region $r_{ut} \sim Multinomial(\psi_{z_{ut}})$
    Draw geo-coordinates $l_{ut} \sim \mathcal{N}(\mu_{r_{ut}}, \Sigma_{r_{ut}})$
    for *each word n in tweet (u,t)* **do**
      Draw category $c_{utn} \sim Multinomial(\pi_{p_{utn}})$
      if $c_{utn} = 0$ **then**
        Draw $w_{utn} \sim Multinomial(\phi^g)$
      else if $c_{utn} = 1$ **then**
        Draw $w_{utn} \sim Multinomial(\phi^a_{z_{ut}, r_{ut}})$
      else if $c_{utn} = 2$ **then**
        Draw $w_{utn} \sim Multinomial(\phi^o_{z_{ut}, r_{ut}})$
    end
  end
end

TABLE I: Notations used in this paper.

Symbol	Description	Symbol	Description
U	number of users	λ	Dirichlet prior vector for π
T	number of tweets by a user	μ	geographical center
N	number of words in each tweet	Σ	regional covariance
K	number of topics	θ	user-topic distribution
R	number of geographical regions	ψ	topic-region distribution
P	number of POS tags	π	POS tag-category distribution
V	the size of vocabulary	ϕ^g	general words distribution
z	topic	ϕ^a	aspect words distribution
r	geographical region	ϕ^o	opinion words distribution
l	location (latitude, longitude)	$m_{p,c}$	number of tokens with POS tag p in category c
w	word	$n^i_{u,t,g}$	number of times tweet t by user u occurs in topic i and region g
c	category (aspect, opinion, general)		
p	POS tag	$l^r_{c,i,g}$	number of times r^{th} word from vocabulary is assigned category c, topic i and region g
α	Dirichlet prior vector for θ	V_{ut}	set of words in tweet (u,t)
β	Dirichlet prior vector for ϕ^a, ϕ^o, ϕ^g	n^{ut}_w	number of occurrences of word w in tweet (u,t)
γ	Dirichlet prior vector for ψ		

B. Model Inference

We use Gibbs-EM algorithm [23], [24] for inference in GASPOP model. We first integrate out the model parameters θ, ψ, π, ϕ^g, ϕ^a, and ϕ^o. After this, the latent parameters left in the model are μ, Σ, z, r, and c.

1) E-step: In the E-step of the inference algorithm, we sample the latent variables: z, r, and c.

Sampling z_{ut} : For each tweet (*u, t*), we first sample the topic z_{ut} as per the following equation:

$$P(z_{ut}=k|*) \propto \frac{n^{k,-ut}_{u,(.),(.)} + \alpha_k}{\sum_{i=1}^{K} n^{i,-ut}_{u,(.),(.)} + \alpha_i} \cdot \frac{n^{k,-ut}_{(.),(.),g} + \gamma_g}{\sum_{r=1}^{R} n^{k,-ut}_{(.),(.),r} + \gamma_r} \cdot \frac{\prod_{w \in V_{ut}} \prod_{j=0}^{n^{ut}_w - 1} (l^{w,-ut}_{0,k,(.)} + \beta_w + j)}{\prod_{j=0}^{n^{ut}_{(.)} - 1} ((\sum_{w=1}^{V} l^{w,-ut}_{0,k,(.)} + \beta_w) + j)} \cdot \frac{\prod_{w \in V_{ut}} \prod_{j=0}^{n^{ut}_w - 1} (l^{w,-ut}_{1,k,g} + \beta_w + j)}{\prod_{j=0}^{n^{ut}_{(.)} - 1} ((\sum_{w=1}^{V} l^{w,-ut}_{1,k,g} + \beta_w) + j)} \cdot \frac{\prod_{w \in V_{ut}} \prod_{j=0}^{n^{ut}_w - 1} (l^{w,-ut}_{2,k,g} + \beta_w + j)}{\prod_{j=0}^{n^{ut}_{(.)} - 1} ((\sum_{w=1}^{V} l^{w,-ut}_{2,k,g} + \beta_w) + j)} \quad (1)$$

Sampling r_{ut} : After sampling the topic z_{ut} for the tweet (*u, t*), we sample its geographical region r_{ut}, conditioned on the topic z_{ut} obtained in Step-1, using Equation(2).

$$P(r_{ut}=g|z_{ut}=k,*) \propto \frac{n^{k,-ut}_{(.),(.),g} + \gamma_g}{\sum_{r=1}^{R} n^{k,-ut}_{(.),(.),r} + \gamma_r} \cdot \frac{1}{\sigma_g} e^{-\frac{1}{2} \frac{(l_{ut}-\mu_g)^T (l_{ut}-\mu_g)}{\sigma_g^2}} \cdot \frac{\prod_{w \in V_{ut}} \prod_{j=0}^{n^{ut}_w - 1} (l^{w,-ut}_{1,k,g} + \beta_w + j)}{\prod_{j=0}^{n^{ut}_{(.)} - 1} ((\sum_{w=1}^{V} l^{w,-ut}_{1,k,g} + \beta_w) + j)} \cdot \frac{\prod_{w \in V_{ut}} \prod_{j=0}^{n^{ut}_w - 1} (l^{w,-ut}_{2,k,g} + \beta_w + j)}{\prod_{j=0}^{n^{ut}_{(.)} - 1} ((\sum_{w=1}^{V} l^{w,-ut}_{2,k,g} + \beta_w) + j)} \quad (2)$$

Sampling c_{utn} : After sampling z_{ut} and r_{ut} for the tweet (*u, t*), we sample the category variable c_{utn} for each word in the tweet, according to the following equations:

$$P(c_{utn}=0|z_{ut}=k, r_{ut}=g, *) \propto \frac{m^{-utn}_{p,0} + \lambda_{p,0}}{\sum_{c=0}^{2} m^{-utn}_{p,c} + \lambda_{p,c}} \cdot \frac{l^{v,-utn}_{0,k,(.)} + \beta_v}{\sum_{r=1}^{V} l^{r,-utn}_{0,k,(.)} + \beta_r} \quad (3)$$

For aspect $(c_{utn}=1)$ and opinion $(c_{utn}=2)$ words,

$$P(c_{utn}=j|z_{ut}=k, r_{ut}=g, *) \propto \frac{m^{-utn}_{p,j} + \lambda_{p,j}}{\sum_{c=0}^{2} m^{-utn}_{p,c} + \lambda_{p,c}} \cdot \frac{l^{v,-utn}_{j,k,g} + \beta_v}{\sum_{r=1}^{V} l^{r,-utn}_{j,k,g} + \beta_r} \quad (4)$$

2) M-step: After sampling the latent variables z, r, and c in the E-step of the inference, we update the regional center μ_r and covariance Σ_r for each region r, in the M-step of the algorithm. Here, s denotes the corresponding Gibbs iteration.

$$\mu_r = \frac{\sum_{s=1}^{S} \sum_{u=1}^{U} \sum_{t=1}^{T} l^{(r_{ut}=r)}_{ut}}{\sum_{s=1}^{S} \sum_{u=1}^{U} \sum_{t=1}^{T} 1^{(r_{ut}=r)}} \quad (5)$$

$$\sigma_r = \sqrt{\frac{\sum_{s=1}^{S} \sum_{u=1}^{U} \sum_{t=1}^{T} (l^{(r_{ut}=r)}_{ut} - \mu_r)^2}{\sum_{s=1}^{S} \sum_{u=1}^{U} \sum_{t=1}^{T} 1^{(r_{ut}=r)}}} \quad (6)$$

TABLE II: λ for POS Tags used in TweetNLP POS Tagger.

P	λ_g	λ_a	λ_o	P	λ_g	λ_a	λ_o	P	λ_g	λ_a	λ_o	P	λ_g	λ_a	λ_o	P	λ_g	λ_a	λ_o
N	15	170	15	O	15	170	15	S	15	170	15	^	15	170	15	Z	15	170	15
L	20	140	40	M	15	170	15	V	20	60	120	A	15	15	170	R	30	30	140
!	15	15	170	D	60	100	40	P	140	20	40	&	170	10	20	T	80	40	80
X	80	40	80	Y	80	40	80	#	170	15	15	@	170	15	15	~	200	0	0
U	200	0	0	E	15	15	170	$	20	140	40	,	200	0	0	G	80	60	60

C. Priors for Model Initialization

The GASPOP model has a bi-variate Normal distribution on the location variable l. The mean μ_r and covariance Σ_r for the regions in R serve as the prior for this Normal distribution. To initialize these parameters, we run the K-means clustering on the tweet geo-coordinates. The values of the mean and average co-variance obtained for the clusters were used as the prior μ_r and Σ_r for latent regions.To estimate the prior λ on the category distribution π, we use a set of human-labeled tweets, with each word labeled with its POS tag p, and the category $c \in \{aspect, opinion, general\}$. The POS tags used were those described in [25]. We then estimate the value of $\lambda_{p,c}$ by calculating the probability $P(c|p)$. The values of λ obtained for the three categories for different POS tags are shown in Table II.

IV. Experiments

A. Dataset Description and Preprocessing

1) Text Data: In order to evaluate the performance of GASPOP in modeling real-world data, we used a Twitter dataset collected using the Twitter Firehose API within a 21-day time interval. This dataset is a 10% random sample of all the tweets that have spatial coordinates and fall within the geographical boundary of the United States. We filtered out all the tweets by users with less than 15 tweets in this time interval, and tweets that had less than 90% English characters. We then normalized the text in the remaining tweets, so that they contained only English characters.

2) Part-of-Speech Tagging and Preprocessing: After text normalization, all the tweets were tagged using the TweetNLP part-of-speech tagger [25], and then preprocessed using basic preprocessing techniques, such as removal of URLs, common stop words, and punctuation marks (excluding emoticons). After this, the dataset contained 2.4 million tweets from 77,482 users, and 1.2 million distinct words.

3) Annotated Data for Evaluation: To quantitatively analyze the ability of our model to classify words as aspect, opinion, or general, based on their POS tags, we prepare a set of 150 manually annotated tweets, with each word annotated as aspect, opinion, or general.

B. Performance Evaluation

1) Evaluation Metrics: For quantitative comparison of GASPOP model against other baseline techniques, we use the following performance metrics:

- Mean Squared Error (MSE): We calculated the MSE using the difference between the actual and predicted location of documents in kilometers using the *haversine formula* [26].
- Perplexity: It is defined as the *negative log likelihood* of test documents using the trained model.
- Precision: For a category c, *precision* is defined as the number of words whose category was predicted correctly among the total number of words predicted for that category.
- Recall: For a category c, *recall* is defined as the number of words whose category was predicted correctly among the total number of words belonging to that category.

2) Baseline Techniques: We compare the performance of GASPOP model against the following baseline models:

- **ME-LDA** [12]**:** ME-LDA aims to model the aspects and opinions from review text. It uses a feature vector, i.e., the POS tag of the previous, current and next word, associated with each word, to discriminate between aspect, opinion and background words.
- **GeoFolk** [19]**:** A spatial topic model that aims to discover the geographical attributes of different topics.
- **Twitter-LDA** [4]**:** This is a topic model that separately models background and topic-related words for Twitter text. After running the model on test documents, we use opinion corpus to classify the topic-related words generated by this model as aspect and opinion.
- A variant of GASPOP model, which does not use POS tags of words in the tweets. We refer to this model as *G-uni*.
- A variant of GASPOP model that has θ as a global corpus-level parameter, instead of a user-level parameter. We refer to this model as *G-global*.

3) Parameter Setting: To initialize GASPOP, the model needs the hyperparameters α, β, γ, and λ as inputs. These hyperparameters serve as the prior information for the model. We used symmetric values for the hyperparameters α, β, and γ, which were derived empirically. Specifically, we set α = 1, β = 0.01, and γ = 5. The values of priors μ, Σ and λ were obtained as discussed in Section III-C.

C. Experimental Results

1) Quantitative Results: In this section, we discuss the quantitative evaluation results of GASPOP model. For empirical evaluation, we ran GASPOP model and its two variants using 100 EM iterations, with 10 Gibbs sampling steps in the E-step of each iteration, varying the number of topics K and the number of regions R. All other baseline models were run using 500 Gibbs sampling iterations.

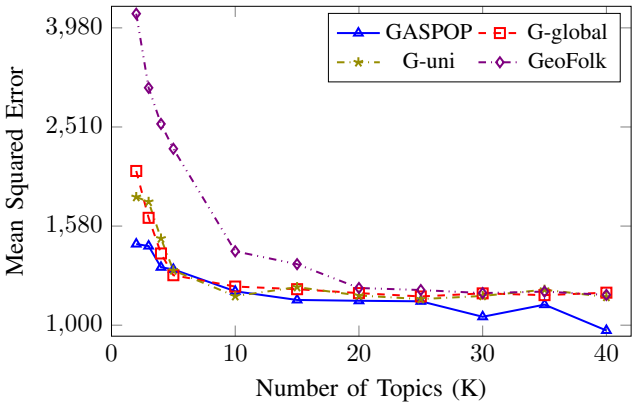

Fig. 2: MSE comparison over different number of regions.

Location Prediction Error: A model with small value of MSE has less location prediction error, and hence better predictive accuracy. We compare the mean squared error of GASPOP model in predicting the location of a tweet, against GeoFolk, and two variants of GASPOP. We first train the models using randomly sampled 90% tweets, and then calculate the MSE for the test dataset containing the remaining 10% tweets. Figure 2 shows the MSE comparison for GASPOP model with 10 topics, by varying R. We can see that GeoFolk model, which does not model topics and regions separately, has the highest MSE among all the models. The main assumption behind the GeoFolk model is that a topic is concentrated only in one geographical region, and hence it calculates the geographical mean and variance associated with topics. This assumption does not hold well in real data, as a common topic like *sports* is equally popular across various regions, and hence the error term will be high for documents related to such topic. We can also see that the MSE for GASPOP is consistently less than its two variants, which shows that user-based modeling in GASPOP, along with syntactic features, such as POS tags, improve the overall predictive accuracy of the model. It can also be seen that the MSE decreases as the number of regions grow. This is due to the fact that samples in a Gaussian distribution converge to the mean of the region. Hence, with increase in the number of regions, the difference between the predicted and actual location decreases.

Perplexity: Since perplexity is the negative log-likelihood, a model with lower perplexity has a better predictive performance. Since we are interested in evaluating the performance of GASPOP model against other location-based models in predicting the topics of documents, we compare the perplexity of GASPOP against its two variants and GeoFolk by varying the number of topics (Figure 3). We can see that the perplexity of GASPOP is consistently less than the baseline models. The perplexity of both GASPOP and G-uni, which have user-topic distribution, is less compared to G-global, which has corpus-level topic distribution. This indicates that, in the case of social media text, the topic corresponding to a document is largely dependent and limited to the interest distribution of the user. It can also be observed that GASPOP has slightly better perplexity than G-uni, which indicates that modeling POS-tags in the generative process improves the overall performance of the model. Also, the perplexity of GASPOP decreases with the number of topics, which is the result of better generalization of dataset by a model with more topics.

Word Category Prediction: To compute the performance of GASPOP in predicting the category of words in unseen documents using a trained model, we compare its weighted-average *precision* and *recall* against the baseline models. Here, in addition to the two variants of GASPOP, we also use ME-LDA, and the variant of Twitter-LDA discussed earlier. For the comparison, we use a test set of manually-labeled tweets, with each word labeled as *aspect*, *opinion*, or *general*. For each of these tweets, we compared the category c of each word obtained after Gibbs sampling iterations, against the human-labeled category, to calculate the precision and recall.

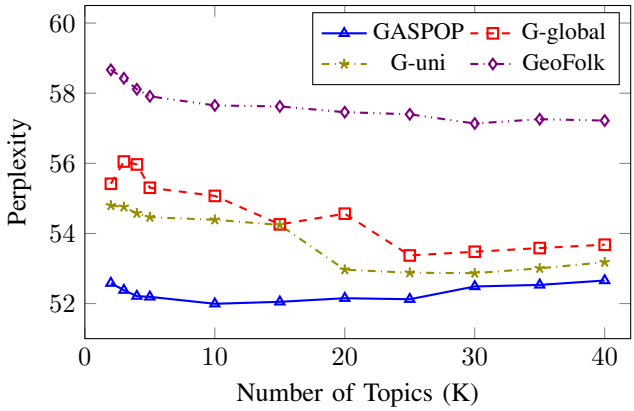

Fig. 3: Perplexity comparison over different number of topics.

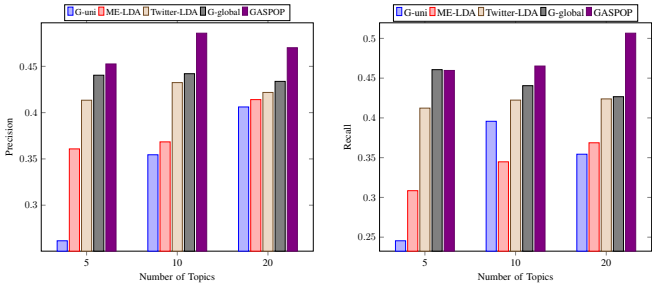

Fig. 4: Precision and recall comparison by varying number of topics.

From the comparison shown in Figure 4, we can see that G-uni, being a fully unsupervised model, has the least precision and recall among all the five models. All the other four models use some supervision for category prediction, and hence perform better in the classification task. It can also be observed that ME-LDA, that uses POS tags of the adjacent words to differentiate between the words from different categories, does not perform very well in the case of Twitter data. Since the text in tweets is highly irregular, model that uses the bag-of-words approach performs better. ME-LDA also uses a multinomial logistic regression function to determine the category of the word. However, in the case of Twitter data, a fully Bayesian model like GASPOP works better because of the irregular structure of the text. Furthermore, the modified version of Twitter-LDA that uses the opinion corpus for classifying words also does not give very good performance, because the opinion corpus does not include the abbreviations that are commonly used in tweets. Using a POS-tagger like TweetNLP, which is specifically designed for such text, helps solve this problem, which is demonstrated by the better performance of GASPOP and G-global among all the models. Finally, we can see that GASPOP model gives the best overall performance amongst all the models, which clearly demonstrates a model that uses bag-of-words approach with the POS tags of words has a significant advantage in modeling microblog data.

2) Qualitative Results: For the qualitative validation of our model, we randomly select one topic, and show its corresponding general words, and aspect and opinion words obtained for four different latent geographical regions, in Table III. It is evident from the words that the topic shown here is related

TABLE III: General, Aspect and Opinion words for the topic *Sports* obtained from different locations.

General Words		
win, play, good, first, watching, hard, start, gold, run, watch		
California	**Aspect**	usa, team, lakers, olympics, dwight, howard
(34.188, -116.795)	**Opinion**	lol, congrats, :), wow, well, omg
Pennsylvania	**Aspect**	nfl, season, preseason, chad, johnson, steelers
(40.675, -80.273))	**Opinion**	haha, #steeltownusa, no, damn, :(, :–
New York	**Aspect**	game, season, team, yankees, york, #mets
(40.766, -74.015)	**Opinion**	lol, :), wow, well, haha, nyc4you
Lousinia	**Aspect**	game, football, lsu, saints, tyrann, mathieu
(32.401, -91.316)	**Opinion**	congrats, well, oh, wow, damn, lmao

to sports. Here, the top-ranked general words are *win*, *play*, *good*, etc. These words are common across all geographical regions for the topic associated with sports. We also present the top-ranked aspect and opinion words for this topic across top four geographical regions, ranked on the basis of topic-region distribution ψ for this topic. The top-ranked region for sports topic has its geographical center close to California. The corresponding top-ranked aspect words from this region are *usa*, *lakers*, and *dwight*. These words correspond to the sports team *Los Angeles Lakers*, which is popular in this geographical region. The opinion words here include *positive-sentiment* words like *lol* and *congrats*, which indicate a positive opinion about this topic, and more specifically, about the sports team Lakers, in this geographical region. Similarly, the other regions also contain words that represent sports teams popular in their respective geographical regions.

V. Conclusion

In this work, we presented GASPOP, a novel probabilistic model to determine latent topics, their aspect and opinion words, from different latent geographical regions, from microblog data. By regarding regions as latent, GASPOP can determine geographical regions based on their lexical features. GASPOP also incorporates the syntactic features of words in the generative process, which helps to classify words as aspect, opinion, or general. The results obtained during the comprehensive evaluation of GASPOP on real Twitter dataset show that our model can discover meaningful results, and outperforms the existing state-of-the-art topic models on widely used evaluation metrics.

VI. Acknowledgements

This work was supported in part by the National Science Foundation grants IIS-1619028 and IIS-1646881.

References

[1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1999, pp. 50–57.

[3] 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.

[4] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *Advances in Information Retrieval*. Springer, 2011, pp. 338–349.

[5] A. Ahuja, W. Wei, and K. M. Carley, "Microblog sentiment topic model," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2016, pp. 1031–1038.

[6] 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.

[7] Z. Xu, R. Lu, L. Xiang, and Q. Yang, "Discovering user interest on twitter with a modified author-topic model," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2011, pp. 422–429.

[8] M. Paul and R. Girju, "A two-dimensional topic-aspect model for discovering multi-faceted topics," *Urbana*, vol. 51, no. 61801, p. 36, 2010.

[9] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 111–120.

[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[11] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Natural language processing and text mining*. Springer, 2007, pp. 9–28.

[12] W. X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 56–65.

[13] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *Language Resources and Evaluation Conference*, vol. 10, 2010, pp. 1320–1326.

[14] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *International AAAI Conference on Weblogs and Social Media*, vol. 10, pp. 178–185, 2010.

[15] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012, pp. 115–120.

[16] R. Wang, W. Huang, W. Chen, T. Wang, and K. Lei, "Asem: Mining aspects and sentiment of events from microblog," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 1923–1926.

[17] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proceedings of the 15th International Conference on World Wide Web*. ACM, 2006, pp. 533–542.

[18] C. Wang, J. Wang, X. Xie, and W.-Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proceedings of the 4th ACM workshop on Geographical information retrieval*. ACM, 2007, pp. 65–70.

[19] S. Sizov, "Geofolk: latent spatial semantics in web 2.0 social media," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 281–290.

[20] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1277–1287.

[21] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 769–778.

[22] A. Kotov, V. Rakesh, E. Agichtein, and C. K. Reddy, "Geographical latent variable models for microblog retrieval," in *European Conference on Information Retrieval*. Springer, 2015, pp. 635–647.

[23] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[24] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[25] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 42–47.

[26] C. C. Robusto, "The cosine-haversine formula," *The American Mathematical Monthly*, vol. 64, no. 1, pp. 38–40, 1957.