Analysis of combining Topic model, Sentiment, Geolocation information approaches on Social Network

Tsung-Hsien Wen [1], *Perng-Hwa Kung* [#2], *Pei-Hao Su* [*3]

Graduate Institute of Electrical Engineering, National Taiwan University
[#] Gomputer Science and Information Engineering, National Taiwan University
[*] Graduate Institute of Communication Engineering, National Taiwan University
[1] r00921030@ntu.edu.tw, [2] r00922048@ntu.edu.tw, [3] r00942135@ntu.edu.tw

Abstract

In this paper, we propose a joint approach of combing Topic model, Sentiment analysis and Geolocation information to discover people's perspective on social network based on these three attribures. Several methods of data preprocessing is adopted to deal with various format of pared data from Twitter. Laten Dirichlet Allocation (LDA) model is used for finding topic distribution within a set of data. Naive Bayes classifier is also adopted to classify senteiments within these data. Furthermore, goelocation information is used for clustering. Experimental results show the effectiveness of our work.

Index Terms: topic model, sentiment analysis, geolocation

1. Introduction

Along with the explosive development of the technology, multimedia and computers play an very important role in our daily lives. Some people tend to post their daily routine and comments on the internet. The field of social network springs up in recent years, such as Facebook, Twitter, Google+. The corresponding analysis of the phenomena in social network also raises rapidly. Studies of social network varies. Some focus on model ing this public phenomenon using mathematical models and theroy [1]. Some focus on the perspective of psychology and human behaviors within this intriguing field [2].

Here we concentrate on using algorithms in the field of computer science and information retrieval. We try to reveal the influence of topic among the masses within specific geolocation. Some related works has been tried to model the relationship of topic and sentiment [3], others focus on the topics within some location [4]. In our work, we try to modeling Topic model, Sentiment analysis and Geolocation information using our joint approach.

2. Senario

In our daily lives, we can find people's comment on some specific TV news. Especially in politics and weather, poeple in different background and location tend to have different opinions. This arouse our interest in finding the relationship and influence of specific topic on some region. Sentiment is often a good indicator to present a person's point of view. Therefore we try to find some massive phenemena on the view of these attributes.

The following sectoins report our approaches and experiments.

3. Approach

In order to discover the joint properties of the three components mentioned, we proposed a modeling framework for it. We will first show the overall framework in Sec. 3.1 to give a overview of our model. Then we will go deeper into each important modules in the following sections.

3.1. Framework

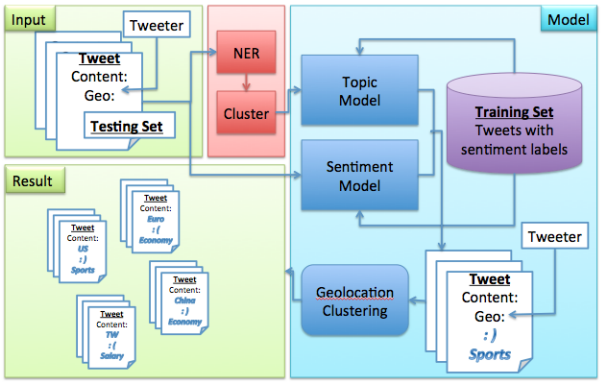

Figure 1: *The overview of modules in our joint sentiment-topic-geolocation model for emotion event discovery.*

Fig. 1 is our model framework for emotion event discovery. Given a set of twits sampled from the Twitter websites, our task is then label the sentiment and the topic for each of them. Before that, we use the NER techniques to extracting import terms as queries and clustering the related tweets together via retrieval. The purpose of do-

ing this as well as the implementation details will be covered in Sec. 3.2. Then the clustered tweets are put into a topic model for topic inference as described in Sec. 3.3. Besides that, a sentiment classifier is also used to classify the emotion for each twit into a binary level, which is discussed in Sec. 3.4. Last but not least, the labeled twits will be clustered according to either its labeled location, the location of the user, or the time zone of the user.

3.2. Name Entity Recognition+Clustering

Recall that Twitter data suffers from two major sources of noises: 1. The 150 character limitation per tweet significantly restricts meaningful contents and introduces a series of Twitter-specific syntatic structures. 2. Users tend to be expressive rather than informative and usually exaggerated contents are prevalent. To effectively identify topics in a diverse set of tweets, it is necessary to filter out redundant information by labeling the words that are the real events, persons, or the like.

To this end, we propose to cluster many different tweets together into aggregated documents for further topic extraction, based on the notion that document clustering provides a good way to group sparse contents in the tweet together [5]. However, note that many conventional document clustering methods take complexity of O($documents^2$), and is also linearly dependent on the number of clusters we try to produce [6]. This makes document clustering difficult to scale up to deal with tens of millions of tweets. Instead, we use name entity recognition technique entailed in [7] to deal with Twitter data by identifying important and co-occuring terms and group the documents accordingly (which is the set of identified name entities). Name Entity Recognition in Twitter corpus is by itself a non-trivial problem, due to the specific syntatic structure introduced by the dataset. The key is to use in-domain data for training Part-of-Speech tagging and recursive structure identification. In this context, recent approaches use linear chained conditional random field as the model. Also, an indepentdent supervised model (here SVM is the used model) is applied to identify whether or not a word's capitalization is actually useful to be included as information. The extracted structure information is fed into another conditional random field model for training name entity segmentation. We simply use the segmentation result (disregarding the name entity labeling model) of name entity recognition as the input for clustering. The entire process is as the following:

1. Train name entity segmentation model $Model_{NE}$ using in-domain and out-domain data
2. For each tweet, collect name entity set $S_i = \{s|\forall s \in tweet_i, isNE(s) = True\}$, where isNE is derived from $Model_{NE}$
3. Determine the total candidate name entity set by combining all name entity sets $NE = \bigcup_i S_i$
4. For each n in NE, set $query = n$, so we have a query set $Query = \bigcup query$, then for all tweets, select $tweets_n = \{tweet|n \in tweet, tweet \in All\_Tweets\}$
5. Build a set of documents with documents defined as $Documents = \{tweet_n|\forall n \in Query\}$

Note that each aggregated document for the new set of documents is a cluster for a particular name entity. After the above process, the resulting clusters of tweets is then fed into Latent Dirichlet Allocation for topic modeling.

3.3. Topic Modeling

Topic model [8], which is based on a statistical generative model that takes into the word co-occurrence relations among documents into consideration to classify the documents into several soft clusters called topics, has been applied on knowledge discovering and latent semantic analysis very succcessfully within a past decode. They have been a popular research topics and several models have been proposed [9, 10, 11]. However, Latent Dirichlet Allocation (aka LDA) [10] is perhaps the most famous and widely used topic model in recent years.

In this work, we propose to model our topic module by conventional Latent Dirichlet Allocation. To model the word generation process, instead of generating a word directly from a document, LDA assumes that behind a set of documents, there's a set of latent variables which is generated from the document to emit the words. The generative process of LDA is shown as following:

1. Draw a topic distribution $\theta^d \sim Dir(\alpha)$
2. For each of the N words w_x in d
   (a) Draw a topic $z_x \sim Multinomial(\theta^d)$
   (b) Draw a word $w_x|z_x \sim Multinomial(\phi^{(z_x)})$, the word distribution over topic $\phi^{(z_x)} \sim Dir(\beta)$

where α and β are Dirichlet hyperparameters for regularization.

Several existing algorithms [10, 12, 13, 14] can be adopted to inference the topic distribution over users θ^d as well as the word distribution over topic $\phi^{(z_x)}$. The collapsed Gibbs Sampler [8, 12] is chosen here for parameter estimation. The collapsed Gibbs Sampling algorithm allows us to compute the joint distribution $P(\vec{w}, \vec{z})$ by integrating out θ^d and ϕ^z and building a Markov Chain whose transition distribution is written as

$$P(z_x = k|\vec{z}_{-x}, \vec{w}) \propto \frac{n_{k,-x}^{(w_x)} + \beta}{\sum_{w \in V} n_{k,-x}^{(w)} + |V|\beta} \cdot \frac{n_{k,-x}^{(d)} + \alpha}{\sum_{k=1}^{K} n_{k,-x}^{(d)} + K\alpha} \quad (1)$$

where k is the index of latent topic, K is the total number of topic classes, $n_{-k}^{(\cdot)}$ is a count that doesn't include the current assignment of z_x, and V is the vocabulary used. Given the Markov chain, one can the approximate the θ and ϕ by

$$\phi_k^{(w)} = \frac{n_k^{(w)} + \beta}{\sum_{w \in V} n_k^{(w)} + |V|\beta} \tag{2}$$

$$\theta_k^{(d)} = \frac{n_k^{(d)} + \alpha}{\sum_{k=1}^{K} n_k^{(d)} + K\alpha} \tag{3}$$

3.4. Sentiment Analysis

Learning the sentiment through texts from people is a well-studied field in past few years. The main purpose of this is to determine public's perspective on specific product or phenemenon. Here we focus on the issue of emotion, that is, positive or negative attitude. Premitive works try to list all possible key words in each emtion, simply search through this database to determine the emotion. This is very intuitive and helpful. However,it is also very costly and ignores the relevance between words. Advanced studies install supervised classifier. Related works such as Naive Bayes, MaxEntropy and SVM, these classifiers are proven useful based on their strong background of mathematical basis. With well-defined feature list, positve and negative data are determined from existing data. These training data are used to train the classifier, it learns the weight of every feature, and is used for classifying new incoming data.

Among the above classifiers, Naive Bayes classifier is adopted in our wok. Each incoming document *d* is assigned to class $c^* = argmax_c P(c|d)$ by Bayes' rule,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \tag{4}$$

where $P(c)$ plays no role in selecting c^*. To estimate the term $P(d|c)$, Naive Bayes decomposes it by assuming that every feature f_i is conditionally independent with others given d classes:

$$P_N B(c|d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)} \tag{5}$$

add-one smoothing is also used to prevent 0-frequency features. Despite its simplicity and the assumption of condition indepencies between each feature, Naive Bayes-based text classification still performs surprisingly well. Some related works also reveal that Naive Bayes is optimal for certain problem with highly dependent features [15].

4. Experiment

In the following section, we setup a experiment for our sentiment-topic-geolocation joint model on Twitter corpus and evaluate its performance based on both quantitative and qualitative analysis. In Sec. 4.1 we talk about how to preprocess the noisy twits. We then move on to our experimental setup in Sec. 4.2. Finally, the results are shown in Sec. 4.3 as well as some discussions and analysis.

4.1. Data Preprocessing

Before we move on to our experiments, it is necessary to preprocess our twitter corpus beforehand since it is quite noisy and multi-layer-oriented in several points of view. For our experimental purpose, we only adopt raw English corpus for topic modeling while we adopt both raw English corpus as well as emoticons for sentiment analysis. The preprocess steps are listed in the Table. 1

Table 1: Data Preprocessing Steps

Target	Description
Tags & Symbols	Use regular expression to remove the unnecessary tags, such as user tags,links location tags, meaningless symbols...etc.
Non-Eng.	Use an English lexicon obtained from Corncob [1] corpus to filter out words that are not in English.
Normal form	Use regular expression to normalize the words such as *goooood* back to its normal form *good*.
Stop word	Consulting a external source of stop word list from internet, filter out those words in the list.
Stemming	Use a Porter Stemmer to stem the words into a unified format.

4.2. Experimental Setup

For the experiments, we use the collected tweets from Twitter as the dataset for model training and testing. Twitter provides its own API for parsing and collecting tweets [2]. The dataset is collected from Nov. 2011 to Dec. 2011, with a total of 40 million tweets. After preprocessing, we are left with roughly six million English tweets. Since the API produces a random sampled stream of tweets, it is not feasible to aggregate social network information. The data types we extracted for the experiment are tweets, user ID, and user's location information.

Regarding out-domain data, we use Corncob dataset [3] to obtain the list of English lexicon. Also, we use GeoWorldMap [4] to obtain a list of city-wise, regional, and country-wise code of the world. The overall data information is summarized in Table .

For implementation, we use the toolkit developed by the authors of to obtain name entity segmentation, and

[2]TwitterAPI: https://dev.twitter.com/

[3]Corncob: http://www.mieliestronk.com/wordlist.html

[4]GeoWorldMap: http://www.geobytes.com/freeservices.htm

we use the MALLET package [16] for LDA training and inference. All experiments are conducted on lab workstation of AMD Opteron 2350 CPU with 16 cores.

Sentiment classifier used is Naive Bayes classifier. In order to obtain true positive and negative emtional tweets, we use list of emoticons. Totally 90000 positive and 220000 negative emotional tweets are extracted to train the classifier. 10-fold cross validation is used to verify the performance. Accuracy of 80.2% is obtained, proven that Naive Bayes is indeed useful in our work. The experiment shown in sectin 4.3.3 is based on the classification of Naive Bayes.

4.3. Experimental Results

Here we present our experiment results. Note that due to time and space of the project, we only have the rough prototype, so we will mainly show qualitative results.

4.3.1. Name Entity Recognition+Clustering

Using the results from name entity recognition, we are able to isolate useful entities and aggregate similar tweets together into a larger document. After deleting stopwords and normalizing, each document becomes a group of related concepts. As shown in Table. 2, we can see that in this case, there are many tweets discussing the football club, Juventu F.C. along with other football leagues.

*juventu team beaten quarter final italian cup lazio alreadi lost game u live arena indosiar wib ar readi prepar
*damn juventu hard transfer market didnt thei sign juventu claudio complet total pass
*o stai unbeaten juventu ar onli team seri juventu back top
*unbeaten juventu return top itali seri win juventu back top
*juventu deni report thei ar interest sign carlo tevez
*ar juventu final back ar steal mate napoli crap seri miss becam european champion juventu napoli
*download napoli juventu game anybodi

Table 2: A sample cluster of tweets, grouped according to "Juventus", a football club in Italy

4.3.2. Topic Modeling

After running LDA on the clustered documents, we arrive at a set of topics. These topics are much more coherent compare to running LDA directly on tweets. Observing Table. 3, we can see that other than some meaningless topics, most topics seem to belong to some particular category. First note that meaningless topics are inevitable and would often be either grammatical terms or everyday life sentence (due to short sentence structure and prevalence of such substituting slangs in a sentence).

Next we discuss some of the meaningful topics. One of the topic type that are prevalent is event-directed topics. For example, the topic "christma season tree holidai santa ipad" indicates Christmas holiday season and gift buying are associated. Another type of topic is functional topic that belongs to a larger category. For example, "game team plai beat footbal state", which belongs to the category "sports" and will likely be the topic of discussion regardless of the time.

tweet topic-most important words	score
*christma season tree holidai santa ipad	0.09083
*happi birthdai famili thanksgiv enjoi dai	0.09553
*black fridai shop store ball sale	0.08126
-school befor high nba chri brown	0.05359
-job ic vote mouth cream fly	0.03948
-parti place open ladi club pm	0.19694
-ar nation countri polic ha bill	0.20649
-song show amaz sing danc voic	0.14024
* event based, - category based	

Table 3: A sample list of topics listing frequent words and corresponding score

Another worthnoting observation is that no geo-location-specific topic is found in the result topics. A possible reason is location-based topics do not become significant amongst all tweets. In addition, the limiting space of tweets usually steer people to say more personal events than location-specific events (e.g. "today's SO HOT, and we're out there marching!"), or say in a personal manner so to make it hard to ascertain exact event.

4.3.3. Combining Results with Sentiment and Geolocation

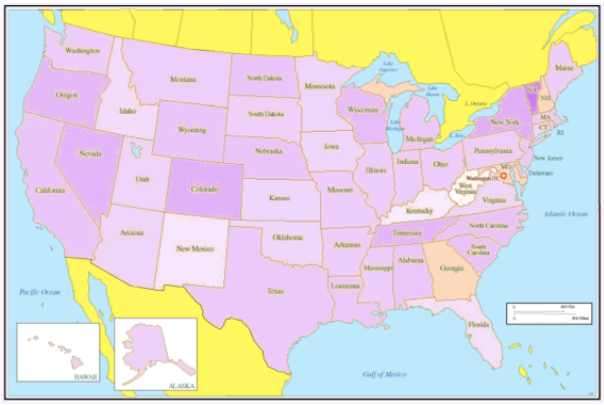

Figure 2: *The sentiment level map for united states, the darker the color indicates the higher the happiness level.*

Using Naive Bayes model, we use the confidence score for each tweet to indicate the strength of the emotion. To see how emotion is distributed in different regions, we choose the United States as the visualizing example. The corresponding emotion map is shown in

Fig. 2. Each state has its own color, with light colors indicate lighter mood and darker colors indicate heavier mood.

For the sake of explanation, we isolate three states for topic-emotion comparison. Consider three states in America: California, Texas, and Colorado. For each state, we generate a score for each topic using the following formula:

$$Score(topic_i) = \Sigma_i emotion(t) * P_t(i), \forall t \; in \; tweets \tag{6}$$

where $P_t(i)$ is the LDA generated probability for topic i of tweet t (emotion is a binary +1/-1 function). The related topic placement can refer to Table. 4 and Table. 5 (we only show CA and TX for brevity). We find that three states share similar topics for positive mood, such as Christmas, food, and idol show. Disparities occurs for some topics: relationship and Black Friday is more stressed in California; football and family more stressed in Colorado and Texas. Regarding negative mood, education is more stressed in Colorado and Texas; voting is more stressed in Texas.

To summarize briefly, we observe that general topics of interest throughout different regions do align with the respective emotion. People usually feel the same way about a topic. Variance exists for certain topics. However, there does not seem to be a location-coherent topic: most feelings toward topics are the strength of spread in related news throughout the web, which coincides with the property of Twitter as a news media.

Positive Topics	Negative Topics
*christma season tree	*job ic vote
holidai santa ipad	mouth cream fly
*song show amaz	*woman dude kiss
sing danc voic	cake man dumb
*black fridai shop	*class teacher alon
store ball sale	studi boyfriend colleg

Table 4: A sample list of positive/negative topics for California ordered from top down

Positive Topics	Negative Topics
*christma season tree	*job ic vote
holidai santa ipad	mouth cream fly
*song show amaz	school befor high
sing danc voic	nba chri brown
* man citi half	class teacher alon
unit goal match	studi boyfriend colleg

Table 5: A sample list of positive/negative topics for Texas ordered from top down

5. Conclusion

In several previous works, both sentiment analysis for text and topic modeling for documents are very popular research domains. Some of them even took the geolocation relations into consideration which gave many interesting analysis of the correlation between them. However, they didn't step a further move to unify topics, sentiments, and geolocation together to give a complete analysis of the three. In this paper, we propose a new framework for jointly model the three modules together in which the NER is used first to extract the name entities as queries to cluster the twits via retrieval, and following up is the topic inference and sentiment labeling for each twits. Finally, the twits are clustered according to its geolocation and yield the experiment results. In our result, it shows that the model has the capability to discover the topics and sentiment relations based on a particular geospace. It gives us the benefit to extract the main cause of certain emotional events on a specific geolocation, which can be an potential information source for the guide of the future political policy direction of governments.

6. References

[1] "Social network analysis: Methods and applications," 1997.

[2] J. Galaskiewicz and S. Wasserman, "Advances in social network analysis: Research in the social and behavioral sciences," 2005.

[3] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," 2009.

[4] J. H. . C. Z. . Z. Y. Zhijun Yin 1, Liangliang Cao 2, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," 2011.

[5] N. O. Andrews and E. A. Fox, "Recent Developments in Document Clustering," Tech. Rep., 2007.

[6] O. Tsur, A. Littman, and A. Rappoport, "Scalable multi stage clustering of tagged micro-messages," in *Proceedings of the 21st international conference companion on World Wide Web*, ser. WWW '12 Companion. ACM, 2012.

[7] A. Ritter, S. Clark, and O. Etzioni, "Named entity recognition in tweets: An experimental study," 2011.

[8] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 5228–5235, 2004.

[9] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '99. ACM, 1999, pp. 50–57.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, ser. UAI '04, 2004.

[12] G. Heinrich, "Parameter estimation for text analysis," Tech. Rep., 2004.

[13] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '08. ACM, 2008, pp. 569–577.

[14] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *NIPS*, 2006.

[15] P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple bayesian classifier," vol. Machine Learning 29:103–130., 1997.

[16] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.