Tracking Drinking Behavior from Twitter Data

COSC 526 Course Project Report

Zhifei Zhang Yajun Wang Yongli Zhu

Department of Electrical Engineering & Computer Science
The University of Tennessee, Knoxville

Abstract

This report presents the work of a tracking drinking behavior form Twitter data set. In this work, the *Latent Dirichlet Allocation* method is adopted with the assistance of human-created key words dictionary. Two main behaviors are extracted: the user device platform usage distribution and user location distribution. From the results, the IOS platform based device got the No.1 place; while the location behavior indicates a similar distribution pattern in accordance with the whole Twitter uses distribution. Those information can be utilized to instruct more effective advertising and more user-oriented apps developing for smart phone/tablet groups.

1 Introduction

Big data is everywhere people look these days. Businesses are falling all over themselves to hire data scientists, privacy advocates are concerned about personal data and control, and technologists and entrepreneurs scramble to find new ways to collect, control and monetize data. It is universally acknowledged that data is powerful and valuable.

Data Mining is an analytic process designed to explore big data in search of consistent patterns or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct applications [1].

Platforms such as Google, Facebook and Twitter are the new factory floor, and online users, who leave digital crumbs as they browse the web and tap into social networks, generate big data that can be bought and sold. Every tweet tweeted, badge unlocked, website searched and “Like” button clicked add to the growing inventory of user information. This rich source of social data is a great point for social data mining because of its inherent openness for public consumption, clean and well-documented API, rich developer tooling, and broad appeal to users from every walk of life. Twitter data is particularly interesting because Tweets happen at the "speed of thought" and are available for consumption as they happen in near real time, represent the broadest cross-section of society at an international level, and are so inherently multifaceted. Tweets and Twitter's "following" mechanism link people in a variety of ways, ranging from short but often meaningful conversational dialogues to interest graphs that connect people and the things that they care about.

One behavior that has the most frequency and periodicity is drinking. Through the social network, people are easy to show their emotions and feelings. Such information as positive or negative opinions about drinking, preference platform and locations of drinkers can be obtained.

Tracking people’s drinking behavior from Twitter data allows government not only to identify effect of certain event on the public but also to take actions to prevent unexpected traffic accident. Furthermore, in a world of endless information sharing, consumers have become the product. Data mining then can be applied to sort it, package it, market it — and companies use it to better target wine customers.

2 Algorithm Description

2.1 Latent Dirichlet Allocation

1) Basic idea

In natural language processing, Latent Dirichlet Allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as a graphical model for topic discovery by David Blei, Andrew Ng, and Michael Jordan [2].

In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis, except that in LDA the topic distribution is assumed to have a ***Dirichlet prior***. In practice, this results in more reasonable mixtures of topics in a document.

For example, an LDA model might have topics that can be classified as CAT_class and DOG_class. A topic has probabilities of generating various words, such as milk, meow, and kitten, which can be classified and interpreted by the viewer as "CAT_class". Naturally, the word cat itself will have high probability given this topic. The DOG_class topic likewise has probabilities of generating each word: puppy, bark, and bone might have high probability. Words without special relevance, such as the “bird”, will have roughly even probability between classes (or can be placed into a separate category).

2) Mathematic model

With plate notation, the dependencies among the many variables can be captured concisely. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. M denotes the number of documents, N the number of words in a document. Thus:

α is the parameter of the Dirichlet prior on the per-document topic distributions,

β is the parameter of the Dirichlet prior on the per-topic word distribution,

θ_i is the topic distribution for document i,

ϕ_k is the word distribution for topic k,

z_{ij} is the topic for the jth word in document i,

w_{ij} is the specific word

The w_{ij} are the only observable variables, and the other variables are latent variables. Mostly, the basic LDA model will be extended to a smoothed version to gain better results. The plate notation is shown in the Fig.1, where K denotes the number of topics considered in the model and:

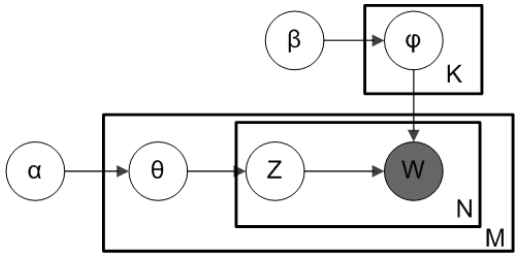

Fig.1 Plate notation representing the LDA model

ϕ is a K*V (V is the dimension of the vocabulary) Markov matrix each row of which

denotes the word distribution of a topic.

The generative process behind is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA assumes the following generative process for a corpus D consisting of M documents each of lengthN_i:

Step1. Choose θ_i ~ Dir (α), where i ϵ{1…M} and Dir (α) is the Dirichlet distribution for parameter α;

Step2. Choose ϕ_k ~ Dir (β), where k ϵ{1…K} and Dir (β) is the Dirichlet distribution for parameter β.

Step3. For each of the word positions i, j, where j ϵ{1…N}, and i ϵ{1…M}

(a) Choose a topic z_{ij} ~ Multinomial(θ_i).

(b) Choose a word w_{ij} ~ Multinomial (ϕ_{ij}).

Note that the Multinomial distribution here refers to the Multinomial with only one trial. It is formally equivalent to the categorical distribution.

2.2 Sentiment analysis

Sentiment is the attitude, opinion or feeling toward something, such as a person, organization, product or location. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and machine learning techniques to identify and extract subjective information in source materials [3].

Sentiment = feelings / Attitudes / Emotions / Opinions。

It is a ***subjective impressions,*** not ***facts.***

For example, some typical scenarios that Sentiment analysis can apply are:

● is this product review positive or negative?

● based on a sample of Tweets, how are people responding to this ad campaign/product release/news item?

● how have bloggers' attitudes about the president changed since the election?

In our project, the drinking behavior Tweets are classified by two sentiment categories: positive and negative. The “positive” sentiment represents the drinker was in happy/excited mood; otherwise, angry/upset/disappointed for “negative” category.

For example, “Hi, guys, I got a job offer today, let’s celebrate it and have a drink!” reflects positive mood of the Twitter user. “Oh my Gosh, my Brazil lost 0-7 to German at our fifth drink round…” indicate he/she was shocked and disappointed by the soccer result, thus belongs to “negative” category.

The sentiment analysis is based on the LDA algorithm and human-assisted pre-classification, i.e., we pre-define a list of highly-representative words dictionary for positive and negative sentiments, then apply this dictionary during the LDA algorithm.

3 Implementation

This section implements the tracking algorithm of drinking behavior. Roughly, the mothed involves three steps: 1) generate corpus for LDA, 2) pick drinking related words from LDA, and 3) extract drinking related Tweets from the raw Twitter data set. The data set we play on is an about 20GB Twitter data, most of which are related to alcohol, but not all of them are tightly concern to drinking. Therefore, the 20GB data set need to be refined into a smaller and more drinking specific data set. Then, a series of analysis, such as sentiment estimation and statistical analysis, can be performed in parallel using the Hadoop.

3.1 Data preparation

The data set we use is Twitter data from March 27th to May 1st in 2014. This is a pre-filtered data set, which is mostly related to alcohol. Total account of the Tweets is around 600 million. When we check into the data, not all Tweets are so related to alcohol or drinking. So, extracting drinking related Tweets is necessary to squeeze valuable parts from the raw data set. The raw Twitter data is stored in zipped JSON format on the distributed file system (DFS) of Hadoop. Noted that it is unnecessary to unzip those zipped JSON files because Hadoop has unzip them automatically. In addition, Hadoop reads those files line by line, thus we need to check the JSON format for each line first and then parse it using the GSON package provided by Google.

3.2 Developing environment

Given the relatively large data set (about 600 million Tweets), we shrink it into a much smaller subset (over 11 million Tweets) by removing those Tweets less related to drinking[4]. This subset focuses more on drinking and yields more reliable analytical results. LDA is applied here to provide drinking related topics, as well as relevant words. However, LDA needs a corpus as initial input, thus we have to go through the raw data set first to collection a corpus that is desired to be as small and representative as possible. In order to obtain more meaningful words, we compare each word from the Tweets with the English word list and unmeaning word list, respectively. The English word list consists of 235,886 English words, and the unmeaning word list stores those words that may not contribute to topic splitting, e.g., “you”, “is”, “on”, “will”, etc. The two word list work like filters, which filter non-English and unmeaning word in the mapper phase and pass those meaningful English words to the reducer phase. Fig.2 demonstrates the process of collecting corpus. The filter settings in reducer phase filters those words with low frequency. Obviously, a word only occurs several times in a huge amount of Tweets contributes little to topic categorization. The corpus eventually obtained involves 17,752 words, all of which are throw into LDA using mapper and reducer again to estimate the hyper parameters [4], namely α, β and γ.

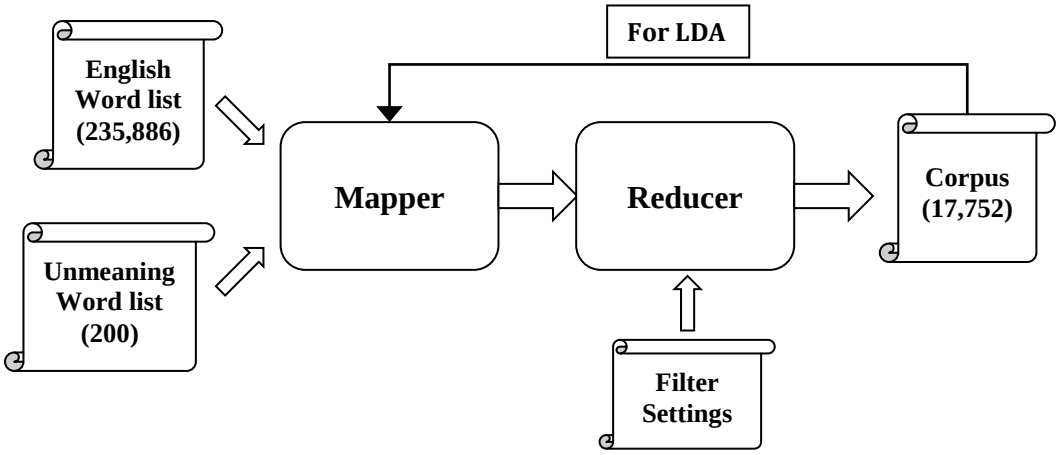

Fig.2 Flowchart of collecting corpus for LDA using Hadoop (mapper and reducer)

Since β describes the relation between topics and words, we pick those word more related to the drinking topic according to β. As list in Fig.3, drunk, wine, hangover and bar are words related to drinking with relatively high probability. Based on the drinking related words, we go through the whole raw data set again to extract the Tweets including one or more drinking related words. At the same time, a sentiment word list is provide to separate positive sentiment from negative ones. All of above is implemented in mapper, and the reducer finally output four sets of data files recording created date of a Tweet, time-zone, source and sentiment. Specifically, <date, cnt> records the count of drinking related Tweets

on each corresponding date. By the same token, <time-zone, cnt> and <source, cnt> record the count of drinking related Tweets for corresponding time-zone and source (platform a Tweet is sent from), respectively.

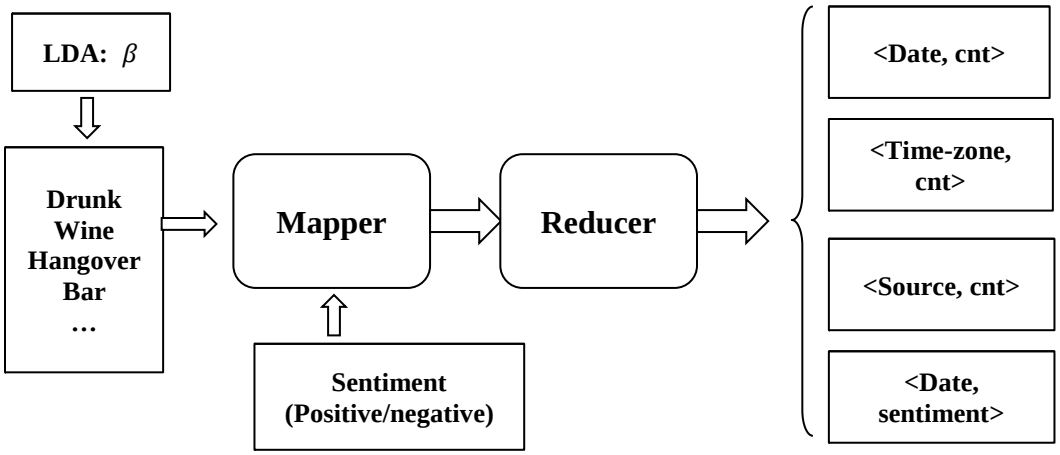

Fig.3 Flowchart of extracting drinking related Tweets using Hadoop (mapper and reducer).

The output data file <date, sentiment> stores the count of Tweets with positive or negative sentiment on each date. Actually, this data file is contracted by two parts---<date, positive sentiment> and <date, negative sentiment>.

In practice, convergence of LDA require tens of iterations which may cost a couple of days. So we only iterated five times and then pick up drinking related word manually based on the hyper parameter β. In addition, the sentiment word list is borrowed from some existing works. In sentiment estimation, both unigram and bigram are employed. Simply speaking, we try to find isolate and adjacent sentiment words in a Tweet. Isolate sentiment word (unigram) refers to a word not connected with any other sentiment word; adjacent sentiment words (bigram) refer to two continuous sentiment words, for example, “don’t like” and “never hate”. The first example is in “negative + positive” format, so it yields a negative sentiment. The second example, however, express a positive sentiment since it is in the format of “negative + negative”. Similarly, the format of “positive + positive” should yield positive sentiment. For unigram, a signal word represent a sentiment.

In a Tweet, it may involves both positive and negative sentiment words. We use a weighted sentiment to give the final sentiment of the Tweet. First of all, sentiment of a Tweet is quantified from -1 to 1, where -1 denotes negative and 1 denote positive. Currently, we just simply set the weight of positive sentiment word as 1 and negative word as -1. If a positive word(s) occur in the Tweet, plus one to the overall sentiment; if a negative word(s) appears, subtract one from the overall sentiment. The final overall sentiment is considered as the sentiment of a Tweet. In a more professional way, the weight of each sentiment word should vary depending on how positive/negative they are.

4 Results and analysis

4.1 Sentiment distribution

The distribution of positive and negative drinking related Tweets are shown in the Fig.4.

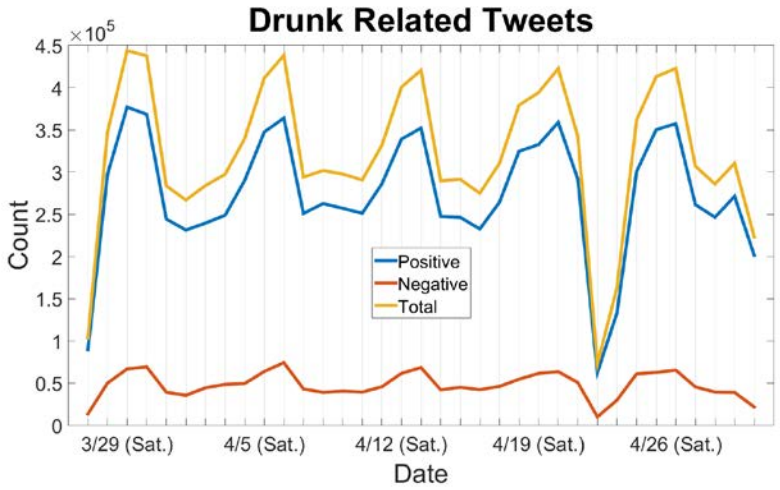

Fig.4 The distribution of positive and negative drinking related Tweets

We can draw three conclusions from this plots:

1) The peaks of drinking –related Tweets happens most frequently during weekend, especially Saturday.
2) Positive Tweets are among the majority of the total drinking related Tweets.
3) There is one abnormal drops between 04/19 and 04/26. At the beginning, we doubt there may be some "abnormal" events happened on that day. But after searching Google for that "abnormal" period, we cannot find any obvious evidence or valuable information related to this "abnormal drop". Finally, after going back to the original data set, we found that the reason is very simple, there is one day missing in the data. Thus, this provides us a "by-product" way to discover any missing data.

4.2 User platform distribution

Fig.5 shows the Source of drunk related Tweets. Nearly half of the drunk related Tweets are sent by IOS platform including iPhone, iPad and MacBook. 18% of drunk related Tweets are sent by Android users. Less than 1% of related Tweets came from Windows Phone. Other kinds of software like Instagram and web page generated almost 30% of the related Tweet.

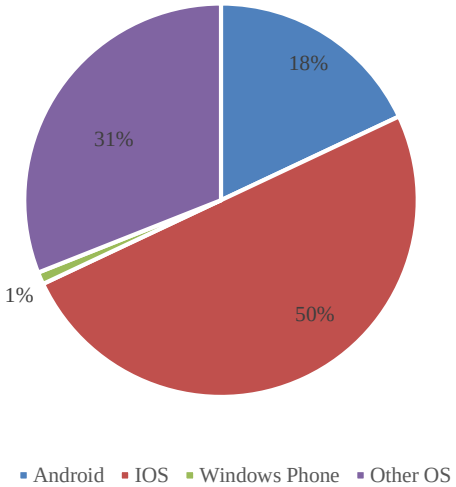

Fig.5 Source of drunk related Tweets

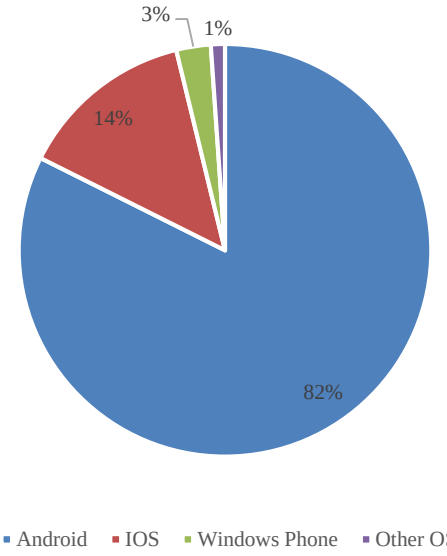

Fig.6 2014 Worldwide Smartphone Market Share (IDC)

In comparison to Fig.6 which shows 2014 Worldwide Smartphone Market Share from International Data Corporation (IDC), the Android platform is the majority component of the markets. The conclusion that people who drink like iPhone more than other brand mobile phones can be draw.

At the same time, several studies and some evidence can support this idea. According to research from venture capital firm Battery Ventures, there is some merit to the idea that iPhones are used by a white collar crowd, while Android favors the blue collar set. Extracting information from a set of survey questions, iPhone users are more likely to have flown in an airplane in the past year, drink wine and have investments in the stock market. Android users, on the other hand, take public transportation, prefer beer, consider themselves religious and have eaten at McDonalds.

Then, the key word “wine” is searched in both App Store and Google Play. There are 2985 Results of “wine” in App Store but only 248 Results of “wine” in Google play. It is obvious that wine or alcohol promotion can be more effective when conducted through people who have iPhones. The same goes to designated driving market.

4.3 Location distribution

The distribution of locations from which Tweets were sent are shown in Fig.7 and Table 1.

Fig.7 Location distribution of drinking related Tweets

Table 1 46 zones over 10,000 drinking related Tweets

Location	Num.	Location	Num.
'Eastern/Time/(US/&/Canada)'	1349324	'Casablanca'	187121
'Central/Time/(US/&/Canada)'	1141305	'Alaska'	129952
'Pacific/Time/(US/&/Canada)'	783803	'Athens'	102358
'London'	583376	'Brasilia'	92126
'Atlantic/Time/(Canada)'	497300	'Beijing'	74283
'Quito'	490727	'Greenland'	52823
'Amsterdam'	325605	'Chennai'	51933
'Arizona'	291043	'Bangkok'	49231
'Mountain/Time/(US/&/Canada)'	234895	'Edinburgh'	44594
'Hawaii'	198108	'Buenos/Aires'	43854
'Dublin'	38960	'Madrid'	31394
'Singapore'	38276	'Kuala/Lumpur'	29112
'Santiago'	37012	'Paris'	27678
'Sydney'	31784	'Pretoria'	26745
'Mexico/City'	26023	'Caracas'	25682

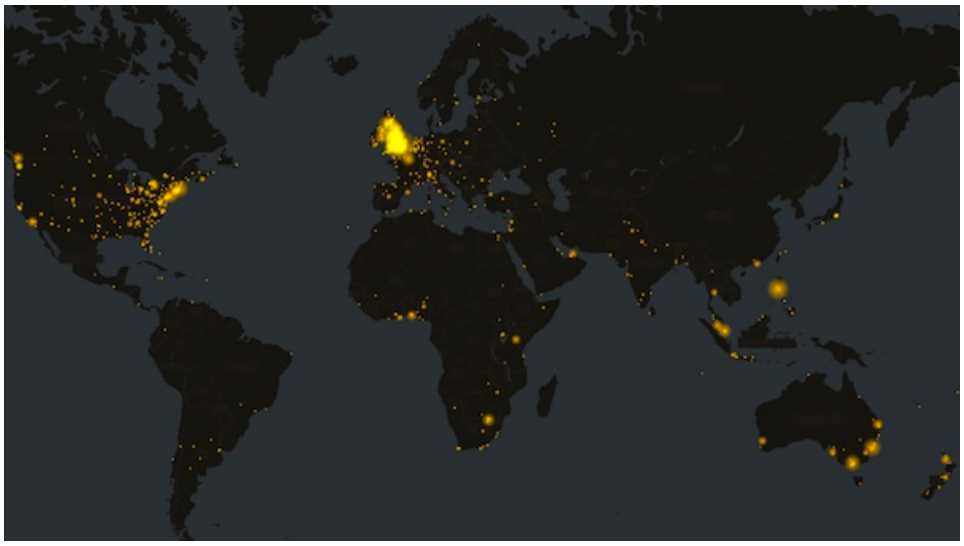

Fig.8 Location distribution of Twitter users

From the Fig.8, one natural conclusion seems to be inferred is, the majority of the alcohol drinkers are from the United States and West Europe. But after serious inspection, we found that this conclusion does not hold rigorously. Because from the Twitter uses distribution (yellow dots), it can be seen that the registered Twitter used located in the U.S. and west Europe. Thus, the blue dots may be not a convincing indicator to support the above conclusion. However, there is still promising application of those location information, for example, the wine manufacturer can produce more wines in advance for some specified location (country/state/province/city) if that place recently has a relatively higher distribution of drinking-related Tweets compared with the remaining area.

4.4 Potential application

Include but not limited to:

- Public safety: e.g. pre-warning to some tourists.
  For example, the government/police department can send warning message/notice to tourists who plan to visit an area with high drinking Tweets distribution.
- Alcohol related business promoting, e.g. giving priority in developing IPhone apps; give priority in advertising investment for those high dinking Tweets areas; increasing stocking amount for those areas, etc.

5 Summary

In this project, the drinking behavior of people is analyzed from Twitter by big data mining technic.

First of all, 60 million Twitter data is obtained. Then LDA and Human method are applied to get the Mapper and Reducer. Third, 11,255,207 drunk related Tweets are used to do the analysis. In the end, drinking behavior of sentiment, time, source and location is presented.

Through big data mining, we may find something hard to be recognized in daily life and identify effect of certain event on the public. At the same time, related marketing can be more targeted. Wine companies may be able to see and predict what their customers like, share, and mention most.

But how to interpret the result of data mining is remained to be answered. Because different interpretation may cause different understanding or even misunderstanding. Another issue is that the extracted behavior features can only represent the pubic trend rather than individual characteristics. Both of them can be left as future study topics.

Acknowledgments

Our team would like to sincerely appreciate the instructor, Dr. Arvind Ramanathan's help and effort in providing us the Tweets data and creating the basic environment of Hadoop in department server for our study convenience.

References

[1] Blei, David M., Ng, Andrew Y., Jordan, Michal I. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022.

[2] Longbing Cao, Philip S. Yu. (2012) Behavior Computing: Modeling, Analysis, Mining and Decision. Springer-Verlag London.

[3] Xiaoran An, Auroop R. Ganguly, Yi Fang, et al. (2014) Tracking Climate Change Opinions from Twitter Data. http://www.cse.scu.edu/~yfang/climate-fang.pdf

[4] Zhai, K., Boyd-Graber, J., Asadi, N., & Alkhouja, M. L. (2012, April). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the 21st international conference on World Wide Web* (pp. 879-888). ACM.