Twitter Sentiment on Tattleware and Bossware: Network Analysis and Topic Modeling Using Latent Dirichlet Allocation (LDA)

Hai Jing(Jane) Tu
tuhai@iu.edu

April 22, 2022

word count= 1,385

1 Introduction

This research aims to use topic modelling to find out more about the sentiment of employee surveillance Tweets collected for paper 2 in ILS-Z639 class. Based on findings from the previous paper, sentiment analysis utilizing Valence Aware Dictionary for sEntiment Reasoning (VADER) shows that there are differences in polarity and intensity levels on the subject of employee surveillance before, during, and near the end of the Covid-19 pandemic. However, as much as we can learn about the compound values that indicate intensity and polarity of the tweets, the result of VADER does not tell us what people actually said about employee surveillance in the tweets.

Using text-processing algorithms to explore text data has been a common practice for social science researchers (Dehghani, Sagae, Sachdeva, and Gratch, 2014,Robinson, Boyd, and Fetterman, 2014, Davidson, Warmsley, Macy, and Weber, 2017, Müller and Guido, 2016). Despite the complexity of natural language, scientists succeeded in developing computer-based approach to capture the emotional, structural, and cognitive components in natural language. Dictionaries often used in text data processing include LIWC (Linguistic Inqurity and Word Count)(Pennebaker, Boyd, and Blackburn, 2015; Robinson et al., 2014) and NLTK (Natural Language Tookkit) (Davidson et al., 2017).

According to Kessel (2018, August 13), Natural language processing (NLP) consists of a large variety of algorithmic methods, both supervised and unsupervised. For example, NLP can use logistic regression, naïve Bayes, random forests, and linear SVMs, etc.(Davidson et al.). Compared with supervised algorithms that require researchers to manually classify data, unsupervised algorithms can find out patterns from documents and summarize the sentiment of the texts without manual classification. Topic modeling uses unsupervised algorithms and allows automatic detection of topic in a corpus of texts when we feed the entire corpus to an algorithm (Müller and Guido, 2016).

Among several topic modeling algorithms that are often used, including non-negative matrix factorization, Structural Topic Models, and Latent Dirichlet Allocation (LDA) (Kessel,2018,

August 13), the LDA model is chosen for this paper. This model tries to find groups of words that occur together frequently and provide insights on what's actually in the Tweets. With LDA, this paper will find the pattern and summarize the content features of the large volume of tweets on the topic of employee surveillance.

In addition to topic modeling, this research also conducts network analysis to find out about the clustering of Twitter users who discuss the topic "employee surveillance." It will help to find out the influential and popular users and how they connect with others in the network.

2 Research Questions

RQ 1. What are the major underlying topics of the historical tweets collected on employee surveillance between 2019 and 2022?

RQ 2. Who are the most influential actors in the twitter network on the subject of employee surveillance?

3 Data

The data for topic modeling are collected using Twitter Application Programming Interface (API). A tweet search query is constructed to include tweets that mentioned "tattleware" or "bossware" or "employee surveillance" between 2019-2022. The query excludes retweets. In total 6,611 tweets are collected.

For this research, all the results from multiple data collection attempts in the previous paper are merged into one dataset for analysis. Also, duplicates are dropped by checking identical tweets from the same twitter id. By dropping duplicates, the sample size drops from 6,611 to 5,451. For the purpose of this research, only the column that contains tweet content is used.

The data for network analysis, on the other hand, are collected using the Twitter Streaming Importer Plugin in Gephi. Using the same search query, a real time network with hundreds or even thousands of nodes and edges can be constructed in Gephi in a short minute.

4 Method

First, Latent Dirichlet Allocation (LDA)(Kapadia (2019, April 14)) is used to analyze the tweets collected on employee surveillance. Second, Gephi will be used to calculate eigenvector centrality and clustering features and visualize a Twitter users network that talks

about employee surveillance.

According to Kapadia,"LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities." By doing topic modelling, we seek to find "topics" that best represent information in a collection of documents.

4.1 Data Preparation

There are different ways to prepare and rescale text data for processing, including bag-of-words representation of text dataMüller and Guido (2016) and TF-IDF (Term frequency-inverse document frequency). For example, Davidson et al. (2017) created weights for bigram, unigram, and trigram features using TF-IDF based on their Twitter data on hate speech.

For this research, bag-of-words are used for text analysis. The 5451 tweets collected using `SNScrape` are loaded into a dataframe for analysis using Python's `pandas` library. Before analysis, the dataset is cleaned by removing punctuations, HTTP links, mentions and hashtags in the tweets. Also, all texts are converted into lowercase. Most importantly, to prepare tweets for LDA analysis, the `stopwords` module is imported from `nltk.corpus` to remove stop words and transform data to serve as input for training in the LDA model.

4.2 Data Analysis

For exploratory analysis, a word cloud analysis is conducted to find out the most frequent words shown in the tweets. The LDA model allows better understanding of individual topics and relationships between the topics. A visualization of the topics and high frequency words will be generated.

According to Müller and Guido(2016), an LDA "topic" is different from what we call a topic in everyday life, as it basically tries to find groups of words that appear together frequently. Not all LDA "topics" make sense. It takes further examination and scrutinizing to identify the most important topics.

Finally, network analysis will reveal the network structure of users who discuss employee surveillance on Twitter. In total 750 nodes are collected. The eigenvector centrality is calculated to indicate the importance of nodes. Using eigenvector centrality as a filter, the majority of nodes with a value under 0.06 are filtered out. The directed edges show how users follow and mention each other.

5 Results

5.1 RQ1

Figure 1 shows the result of a word cloud image of all the tweets created using the `WordCloud` module in `python`.

Figure 1: Word Cloud Based on 5451 Tweets on Employee Surveillance

Using the `gensim` module, a relatively new development is text processing (Müller and Guido, 2016), LDA generates 10 topics where each topic is a combination of keywords and each keyword contributes a certain weight to the topic. Figure 2 shows the 10 topics generated from the analysis.

With the help of `pyLDAvis`, the result of topic modeling can be visualized in an interactive chart(Figure 3). There are three clusters of the top 10 topics.

5.2 RQ2

Figure 4 shows the real time Twitter user network based on the keyword "tattleware" or "bossware" in April 2022. The network is rather loose and most users do not have connection with each other.

To take a closer look at those who do interact with others, I filtered out nodes with less than 1 degree in Gephi, and focused on those who are actually connected in this network (Figure 5). It shows the most influential users in this network that discusses the topic of employee surveillance. Most of the connections are weak ties, only a small number of them form strong ties by having mutual connections with a third node.

```
[(0,
  '0.027*"bossware" + 0.023*"surveillance" + 0.020*"employee" + '
  '0.019*"workers" + 0.012*"invasive" + 0.012*"tracking" + 0.012*"employees" + '
  '0.011*"secretive" + 0.008*"inside" + 0.008*"video"'),
 (1,
  '0.026*"surveillance" + 0.021*"employee" + 0.015*"work" + 0.014*"remote" + '
  '0.012*"tattleware" + 0.012*"app" + 0.008*"time" + 0.008*"employees" + '
  '0.008*"new" + 0.007*"software"'),
 (2,
  '0.050*"tattleware" + 0.042*"home" + 0.035*"employees" + 0.033*"working" + '
  '0.023*"bosses" + 0.019*"tabs" + 0.019*"keep" + 0.019*"turn" + '
  '0.018*"surveillance" + 0.012*"employee"'),
 (3,
  '0.023*"surveillance" + 0.022*"tattleware" + 0.021*"employee" + '
  '0.019*"employers" + 0.015*"track" + 0.013*"home" + 0.012*"china" + '
  '0.011*"time" + 0.010*"spying" + 0.010*"instead"'),
 (4,
  '0.040*"surveillance" + 0.027*"employee" + 0.013*"employees" + '
  '0.011*"bossware" + 0.011*"work" + 0.009*"home" + 0.008*"workers" + '
  '0.007*"amp" + 0.007*"software" + 0.006*"tattleware"'),
 (5,
  '0.018*"surveillance" + 0.017*"using" + 0.012*"employees" + 0.011*"app" + '
  '0.011*"employee" + 0.011*"employer" + 0.011*"called" + 0.010*"viual" + '
  '0.010*"software" + 0.010*"created"'),
 (6,
  '0.025*"app" + 0.020*"phone" + 0.016*"download" + 0.016*"location" + '
  '0.014*"got" + 0.013*"refuses" + 0.013*"monitors" + 0.013*"fired" + '
  '0.012*"school" + 0.012*"custodian"'),
 (7,
  '0.047*"tattleware" + 0.043*"employees" + 0.040*"home" + 0.036*"working" + '
  '0.033*"bosses" + 0.031*"keep" + 0.031*"turn" + 0.030*"tabs" + '
  '0.022*"surveillance" + 0.012*"employee"'),
 (8,
  '0.023*"location" + 0.022*"surveillance" + 0.022*"app" + 0.021*"fired" + '
  '0.021*"custodian" + 0.021*"school" + 0.020*"says" + 0.019*"phone" + '
  '0.019*"download" + 0.018*"got"'),
 (9,
  '0.022*"workers" + 0.022*"surveillance" + 0.018*"bossware" + '
  '0.014*"employee" + 0.013*"tattleware" + 0.013*"invasive" + '
  '0.012*"employees" + 0.011*"secretive" + 0.011*"tracking" + 0.010*"work"')]
```

Figure 2: Top 10 Topics Generated by Topic Modeling using LDA

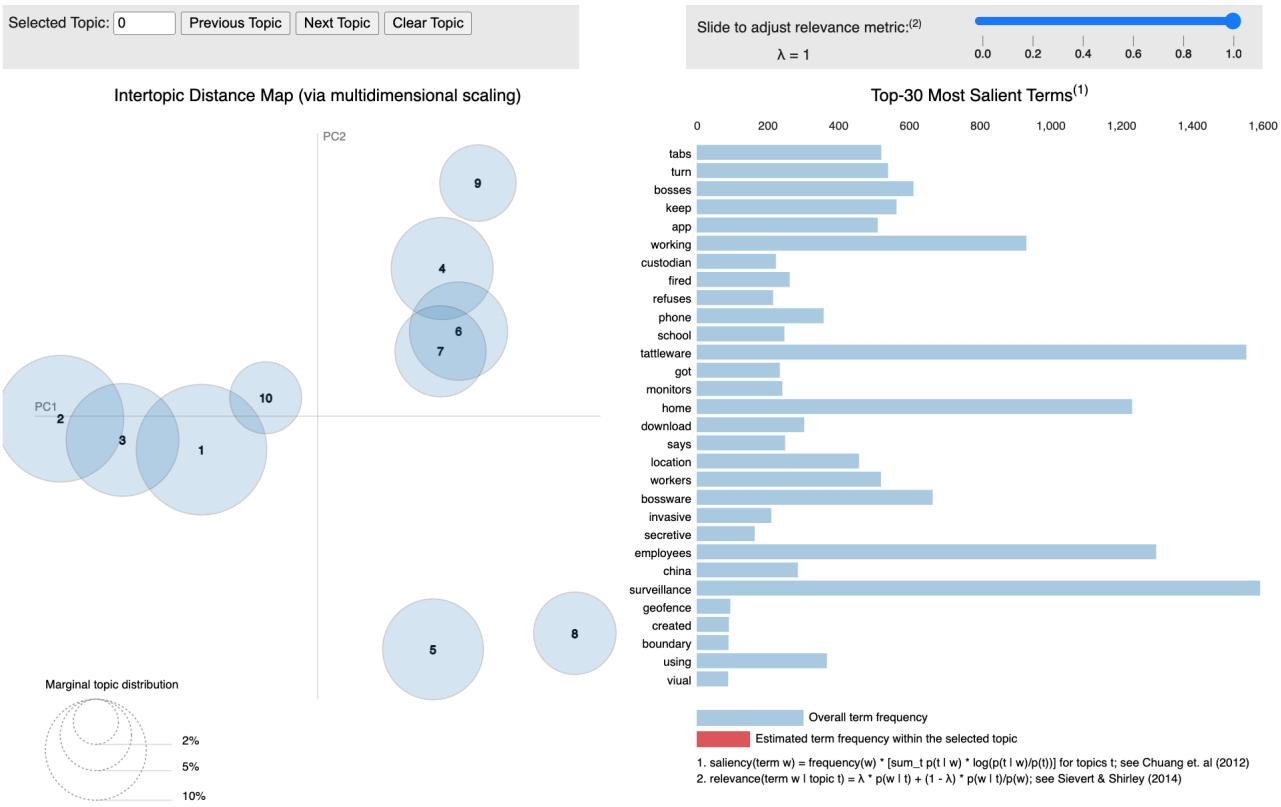

Figure 3: Visualization of Top 10 Topics Using LDA

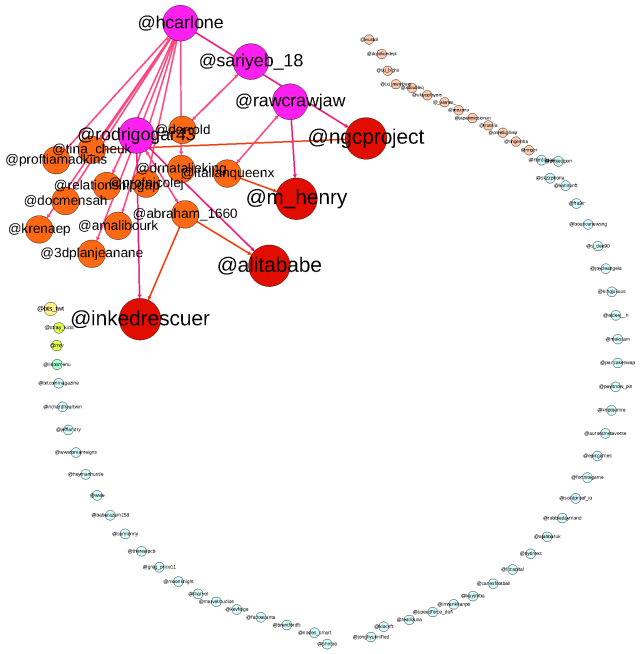

Figure 4: Network Visualization of Twitter Users on the Subject of Employee Surveillance

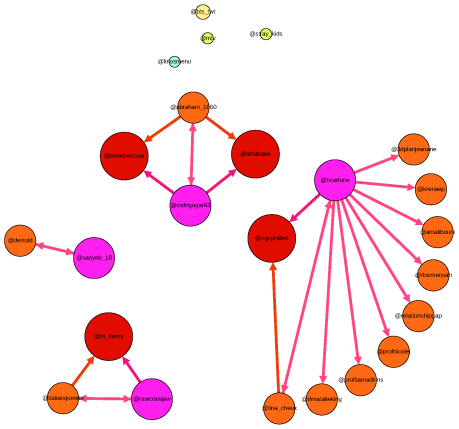

Figure 5: Network Visualization of Twitter Users on the Subject of Employee Surveillance

6 Discussion

This research implements topic modelling and network analysis to find out more about the sentiment on employee surveillance on Twitter. First, the top 10 topics generated by LDA shows ”invasive,” ”secretive” are significant part of the top 1 topic. Other words that are distinctive in the modeling are ”home”, ”remote”, ”fired”, etc. This method definitely allows us to get a better sens of what's included in the collected tweets. But I do notice some irregularities in the topics, such as the frequent recurrences of ”employee” and ”workers.” The insight provided by this topic modeling is still limited. The next step is to use topic model for categorization of these tweets so comparisons between different times can be made.

The network analysis conducted in Gephi based on popularity of nodes measured by eigenvector centrality. The average clustering coefficiency is extremely low (0.072), which means this is a very scattered network. Based on a very limited sample, obtaining a dense network is not likely. But it does show the topic is being discussed by individuals without common connections and may not receive influence from each other. it's reasonable to assume that the discussion on the topic of employee surveillance is rather voluntary than occurring under the influence by opinion leaders.

There are certainly limitations to this research. The full potential of topic modeling is yet to be explored. By experimenting different parameters, there may be more accurate content features being uncovered. Also the sample for network analysis is rather small based on the real time streaming importer. It would be extremely helpful to analyze the twitter users collected in the 5451 tweets collected for this paper to see how they are or are not connected to each other.

References

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512-515. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14955

Dehghani, M., Sagae, K., Sachdeva, S., & Gratch, J. (2014). Analyzing political rhetoric in conservative and liberal weblogs related to the construction of the "ground zero mosque". *Journal of Information Technology & Politics*, *11*(1), 1-14. Retrieved from https://doi.org/10.1080/19331681.2013.826613 doi: 10.1080/19331681.2013.826613

Kapadia, S. (2019, April 14). Topic modeling in python: Latent dirichlet allocation (lda): How to get started with topic modeling using lda in python. Retrieved from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

Kessel, P. v. (2018, August 13). An intro to topic models for text analysis. Retrieved from https://medium.com/pew-research-center-decoded/an-intro-to-topic-models-for-text-analysis-de5aa3e72bdb

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python: A guide for data scientists*. O'Reilly Media.

Pennebaker, J. W., Boyd, R. L., & Blackburn, K. (2015). *The development and psychometric properties of liwc2015*. Autin, TX: University of Texas at Austin. Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

Robinson, M. D., Boyd, R. L., & Fetterman, A. K. (2014). An emotional signature of political ideology:evidence from two linguistic content-coding studies. *Personality and Individual Differences*, *71*, 98-102. Retrieved from https://doi.org/10.1016/j.paid.2014.07.039