Journal of Advances in Mathematics and Computer Science

25(6): 1-21, 2017; Article no.JAMCS.38582

ISSN: 2456-9968

(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Estimation of Community Views on Criminal Justice a Statistical Document Analysis Approach

Sujeong Seo[1*] and Ernest Fokoue[2]

[1]*School of Mathematical Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA.*
[2]*Faculty of School of Mathematical Sciences, College of Science, Rochester Institute of Technology, Rochester, NY 14623, USA.*

Authors' contributions

This work was carried out in collaboration between both authors. Author SS designed the study, performed the statistical analysis, wrote the protocol, and wrote the first draft of the manuscript. Author EF managed the analyses of the study and the literature searches. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2017/38582
Editor(s):
(1) H. M. Srivastava, Professor, Department of Mathematics and Statistics, University of Victoria, Canada.
Reviewers:
(1) Zlatin Zlatev, Trakia University, Bulgaria.
(2) Pasupuleti Venkata Siva Kumar, VNR Vignana Jyothi Institute of Engineering and Technology, India.
(3) Jackson Akpojaro, Samuel Adegboyega University, Nigeria.
Complete Peer review History: http://www.sciencedomain.org/review-history/22537

Original Research Article

Received: 1st December 2017
Accepted: 23rd December 2017
Published: 1st January 2018

Abstract

The Community Views on Criminal Justice System (CVCJS) initiative was established to collect a city community's perceptions on experiences with local Police Departments and other agencies in the criminal justice system, and share those findings to inform local Gun Involved Violence Elimination (GIVE) strategies in New York State. This paper reviews those findings via an empirical study with major text mining methods. Specifically, atomic/canonical words along with as n-grams are used to explore such text mining tasks as sentiment analysis, document clustering and topic modeling, all aimed at gaining insights into all the patterns underlying the community's perception of policing and criminal justice. We use Latent Dirichlet Allocation [LDA] analysis and Structural Topic Model [STM] analysis, which are currently among the most widely used topic modelling algorithms in the fields of computer science, statistics, and machine

**Corresponding author: E-mail: ss1526@rit.edu*

learning. Despite the very limited amount of data available for our study, the combination of sentiment analysis with document clustering and topic modelling helps extract and reveal very interesting patterns underlying the community's views of policing and criminal justice.

Keywords: Text mining; document clustering; sentiment analysis; topic modeling; n-gram; criminal justice; data science; statistical analysis.

1 Introduction

The criminal justice system consists of three major parts: law enforcement, such as local police and sheriff, adjudication, such as courts, and corrections such as jails, prisons, probation and parole. In the United States, criminal justice policy has been guided by the 1967 President's Commission on Law Enforcement and Administration of Justice, which issued a ground-breaking report "The Challenge of Crime in a Free Society". This report made more than 200 recommendations as part of a comprehensive approach toward the prevention and fighting with crime. The Commission justified a set of systems outlining an approach to criminal justice, with improved coordination among law enforcement, courts, and correctional agencies [1]. The President's Commission defined the criminal justice system as the means for society to "enforce the standards of conduct necessary to protect individuals and the community" [2].

In both the literature and the field of law enforcement, the system of allocating police officers to a particular area to build partnerships inside the community already existed, and it has been called community policing. Collaboration between the law enforcement agency, individuals and the organization they serve, help to improve solutions to problems and evaluate trust in police [3]. However, its philosophy, mechanism, and efficiency has been debated [4].

With the recent incidents involving police use of force and other issues, the legality of police action has been questioned in many communities. A large scale of demonstrations and protest marches arose in many cities in 2014 and 2015. Even riots against perceptions of police misconduct and excessive use of force appeared. These conflicts affected community-police relationships, causing many to be untenable. Through this social movement, the importance of police-community relationships has become a major issue in the United States [5]. It brought attention from local law enforcement, to federal agencies, to detect community views on the criminal justice system.

Simultaneously, interest in data science in the Criminal Justice field has been rapidly increasing, such as detection and prediction in cybercrime profiling [6] and criminal activities [7]. Many Criminal Justice research projects are mainly concentrated around qualitative research. With a large amount of text files, researchers can be easily confused with what they can do with the data. Besides, qualitative research, while incredibly useful in this field, often suffers from the limitation of lack of objective measures that are typically the foundation of quantitative analysis. The closest form of quantitative measurement in this context often takes the form of Likert measurements arising from questionnaires. Due to limitations inherent in questionnaires, like the constrained range of possible answers, many researchers who traditionally used either qualitative analysis or questionnaires, have recently been adopting quantitative text analytics, with the anticipated hope of gaining far more insights from respondents' by quantifying their free-form text responses.

Text mining, especially statistical text mining is a broad framework with several methods such as latent semantic indexing (LSI), topic modelling, document classification, document clustering, and sentiment analysis, that allow the quantitative measurements on documents, thereby providing a rich and powerful alternative to previous methods. Authors like [8] have given detailed accounts of

basic text mining. Throughout this paper, the questionnaire portion of the surveys was removed, and only the free-form text was maintained and analyzed. Consider for instance the following three fragments (portions) of the respondents' text from our corpus:

Question 1 - 3:

What types of interactions are you thinking about when you were answering?

In the last six months mostly meetings. Before the last six months, calls made to police

Mostly meetings, my wife and I called the police. We have a lending library in our neighborhood and there was a guy who was throwing books around, pretty minor but they responded quickly.

What makes it good or bad?

I've had interactions with cops who think that because I'm a woman I'm a bad driver. Or other times you know they ask "Do you know why I pulled you over?" which is protocol, but if you don't answer the right way they become somewhat sarcastic with you, and I don't think that the sarcasm is necessary. I would evaluate that as a bad interaction. If I did something wrong tell me straight up. When they are sarcastic it puts people in a sour mood.

Black while driving, people having to tell their black son or daughter how to properly interact with the police, something I wouldn't have to do with my children

Cops come in with presumptions, sure it is implicit bias, but it is just unnecessary. I've been in the car with my friend who is Hispanic but the cop let her go because she is an attractive Latino woman. Even though she was speeding, and she didn't have her license on her at the time.

Would you consider that a bad interaction?

Yes! She should have gotten a ticket. Everyone should get the same treatment regardless of their race/ethnicity, gender, age etc.

My interactions have been somewhat neutral. However, one time I was with my my friend who was having a mental health issue and who also happened to be transgender, the police handled that situation very well. And I called and complemented the officers because I thought it was handled really well. At the same time, I hear about immigrants through my work who have issues with police, black and brown people, trans, LGBQ people who have had various issues with the police and I think that is just not right.

I *hear* more about bad interactions, but I personally have had good interactions so I am neutral. Plus, when answering I was thinking a lot about community meetings through my work which are generally neutral interactions

Our goals with the dataset from CVCJS are (a) to extract, wherever possible, the general sentiment of the community toward the criminal justice system in the city, (b) find out if there are different groups of perception based on document clustering, and (c) attempt to confirm or discover some of the latent structures responsible (or at least related) to the views/perceptions held by the community by via topic modeling. There are several challenges with the text data from CVCJS: very limited amount of data, initial preprocessing reveals very very noisy data due to educational level of respondents, and various levels of note taking skills, meaning that several note takers wrote the documents with their own abbreviation and comments about the nuance of respondents. Through this paper, document analysis assumes that (a) each respondent's text is the document (note that some documents are much longer than others); (b) within the so-called bag of words (BOW) assumption/approach, each document is represented by the words only, without the semantics being taken into account; (c) to avoid instances of missing negations and other compounds, n-grams are used. For instance, trust and not-trust convey two opposite sentiments. By just getting rid of the stop word 'no', this would be lost/missed. Hence the use of n-grams, at least 2-gram; (d) removal of stop words such as 'she', 'he', 'we', 'they', and 'I'. Using the BOW assumption combined with suitable uses of n-grams, our basic data structure after pre-processing, is the so-called term document matrix (tdm) also known as the document term matrix (dtm), which can be written in the following $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & \cdots & \cdots & \cdots & X_{1j} & \cdots & X_{1p} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{i1} & X_{i2} & \cdots & \cdots & \cdots & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \vdots & \ddots & \ddots & \cdots & \cdots & \cdots & \cdots & \vdots \\ X_{n1} & X_{n2} & \cdots & \cdots & \cdots & \cdots & X_{nj} & \cdots & X_{np} \end{bmatrix} \tag{1.1}$$

Each column X_j of $\mathbf{X}$ represents either a atomic word like *police*, *trust*, *arrest* or an n-gram like *not-good*, *dont-trust*, *get-pulled*. In most document analysis tasks, the term document matrix $\mathbf{X}$ is typically very sparse, with 95% of zeroes not unusual. Besides, except in rare cases, $\mathbf{X}$ tends to be

ultra-high dimensional, meaning that $p \ggg n$ as depicted in the matrix, since the number of words tends be much much higher than the number of documents to be text-analyzed. Depending on the analysis, the entries X_{ij} of $\mathbf{X}$ can be of one of the following types:

- $X_{ij} \in \{0, 1\}$, if $\mathbf{X}$ is simply a binary incidence matrix with entry $X_{ij} = 1$ if word j appears in document i, and $X_{ij} = 0$ otherwise.
- $X_{ij} \equiv$ Frequency of word j in document i.
- $X_{ij} \equiv$ logarithmized relative frequency of word j in document i.

As indicated earlier, one of the most natural questions one may seek to answer in the presence of a collection of documents dealing with the views of a community is the following: *are there meaningful and clearly distinct groups in the community with different aspects of the perception of policing? If so, how many groups are there?* As we shall see later we will tackle this question using methods like *Partitioning Around Medoids (PAM)* and *Hierarchical Clustering*. Specifically, if we anticipate k groups of views in the community, and denote by $P_k = C_1 \cup \cdots \cup C_k$, the partitioning of the data into k groups/clusters, then we seek the optimum clustering

$$P_k^* = \underset{P_k}{\texttt{argmin}} \left\{ \sum_{j=1}^{k} \sum_{i=1}^{n} z_{ij} d(\mathbf{x}_i, \mathbf{x}_j^*) \right\}, \tag{1.2}$$

where $z_{ij} = \mathbb{1}(\mathbf{x}_i \in C_j)$ and $d(\cdot)$ can be any distance like the Euclidean $d(\mathbf{x}_i, \mathbf{x}_j^*) = \|\mathbf{x}_i - \mathbf{x}_j^*\|^2$ or the Manhattan distance $d(\mathbf{x}_i, \mathbf{x}_j^*) = \|\mathbf{x}_i - \mathbf{x}_j^*\|_1$, or any other suitable distance. Section 4 of this paper is dedicated to the exploration of the clustering of the documents in our corpus. The other question that naturally arises from such a corpus of documents is: *What are the topics underlying the responses given by the members of the community?* The fundamental idea here is that the documents are just the manifestations of latent constructs/concepts which we will refer to as topics. In words, the density function (mass function) $p(w)$ of a word w is truly a marginal density obtained from the joint density of the word w and the topic z that generate it, specifically

$$p(w) = \int_{\mathcal{Z}} p(w|z)p(z)dz. \tag{1.3}$$

We provide a thorough topic modelling analysis of the data in section 4, right after some general text minind and text analytics definitions and concepts.

2 Generalities of Text Mining and Text Analytics

Generally, text mining is considered a multidisciplinary area of study consisting of data mining, linguistics, computational statistics, computer science, library and information science, and medical science. Latent corpus analysis, text clustering, document summarization, ontology and taxonomy creation, text classification, and categorization, visualization, database technology, machine learning, and data mining are known as standard techniques [9][10].

There are two major approaches in text mining; using linguistic information in both vocabulary and signals through word order, and the bag-of-words model in which only vocabulary affects word order signals, which are all permutations of the words in a document that can provide corresponding data. Although bag-of-words models have limitations such as synonymy and polysemy, they have achieved rich theory and are successful in the context of topic modeling. Linguistic information is correlated to speech recognition and natural language processing, which have very fashionable literatures [8][11].

In the criminal justice field, especially for cybercrime profiling, collaboration with social networking services (e.g. Facebook, Twitter, Snapchat, and so on) has encouraged reporting. Since data mining

allows researchers to explore and analyze large quantities of data, catching significant results such as meaningful patterns, up to date, all the national security organizations rely on data and text mining techniques to determine and forecast criminal activities. Text mining or text data mining is an analysis to extract non-trivial patterns or knowledge from unstructured text. [6][12].

Full semantic analysis with text data is absurd. Even humans often miss nuances and implicit meaning. However, in the natural language processing models, small achievements are important because of its complexity. One major example is Chomskian deep structure. Humans are hardwired for language in Chomsky's model, with deep structure which specific syntax and surface structure that allows that distinguish, say, and verb placement in English from verb placement in German. Text mining through a natural language model of deep or surface structure is roughly the same as the problem of artificial intelligence. This difficulty requires make simple analytical strategies including n-grams [8].

Tokens (or word stems) are the base of most syntactic text mining strategies. Tokens can be individual words, phrases or sometimes whole sentences. The act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols, and other elements is called tokenization. A token is a base word such as 'call', 'called', and 'calls' are identified by a stemmer as being the same. Since stemming permits to reduce the number of words (features) by ignoring tense, pluralization, hyphenation, and other subtleties, it is statistically applicable.

In English, long sequences of words have noticeable semantic signals. It means that if one is given the first word (token) in an n-gram, then the following seven or eight words have probabilities, correlated with the first word and one can measure differences from their overall frequency, but words after this following word, like the ninth word, may not have conditional probability different than its background frequency. According to Laurence Doyle, a communications researcher working with SETI, the Search for Extraterrestrial Intelligence in California, up to nine words away, there are conditional probabilities which are imposed by the rule structures of human languages. [13]

Text mining applications, especially topic models use n-grams in various ways (see references for detail), but it is not a common tool used in the field of Criminal Justice. The major problem with n-grams in text mining is the inferential step. Despite attempts to use learning automatically from training data or designed experiments, some people use human intelligence in the terms of "privileged information".[14] With the high influence of humans in Criminal Justice research, n-grams help to avoid human intelligence bias and preference and find other potential predictors among n-grams. Also, n-grams will help to detect words including negative meaning with linked words. For instance, words "safe" and "unsafe" are easily detected, but if we use simple unigram, we cannot detect words or tokens such as "does not feel safe" or "not safe".

3 Topic Modeling

Blei et al [15] proposed Latent Dirichlet Allocation (LDA) and it became the most popular topic models methodology. LDA selects topics for each document according to one Dirichlet distribution, then conditional on a topic, the vocabulary is chosen according to its corresponding Dirichlet distribution. The generative model for assigning words to documents is

- $\phi_k \in \mathbb{R}$ can be drawn from a Dirichlet distribution with parameter α for each topic; for topic k, it randomly determines the distribution over the vocabulary.
- $\theta_j \in \mathbb{R}^K$ can be drawn from a Dirichlet distribution with parameter β for document D_j which randomly demonstrates the extent to which document D_j participates in each of the K topics.

- For each word in document D_j, first we can draw a single topic z_i from the one-trial multinomial with parameter θ_j, independently. Once it draws the ith topic, then the word is chosen as a single draw from the one-trial multinomial with parameter $\boldsymbol{\phi}_i$.

The following plate diagram is often used to describe this generative model, which represents the relationships between the mechanisms for composing documents as random mixtures of topics with independently drawn vocabulary, with probabilities depending upon the topic [16].

The general definition of topic modeling is that it is a type of statistical model for discovering the abstract, which are topics that occur in a collection of documents. Topic modeling is used to solve the problem of automatically classifying sets of documents into themes. In the machine learning and natural language process, topic models provide a probabilistic framework for the frequency of term appearances in the documents of a given corpus. Using only information related to the frequency of the terms in the text documents seems to ignore their contexts. In natural language processing, topic models extend and build on classical methods such as the unigram model and the mixture of unigram models (Nigam et al. [17]) as well as Latent Semantic Analysis (LSA; Deerwester et al. [18]). Topic models are mixed-membership models so that they are different with the simple unigram or the mixture of unigram models. A unigram is created from n-gram methodology. An n-gram in the fields of computational linguistics and probability is a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words, or base pairs according to the application. The n-grams are typically collected from a text or speech corpus (Broder et al. [19]). An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n-1) order Markov model.

In our model, our basic item is a word. In the unigram model, each word is assumed to be drawn from the same term distribution, but in mixed-membership models documents are not assumed to belong to a single topic. However, it simultaneously belongs to several topics and the topic distributions are vary over documents.

Hofmann [20] proposed the early topic model with probabilistic LSA. He assumed that the interdependence between words in a document can be explained by the latent topics of the document that it belongs to. Conditions of the word appearances in a document in the topic assignments of the words are independent. The latent Dirichlet allocation (LDA; Blei et al. [21]) model is a three-level hierarchical Bayesian mixture model for discrete data where topics are not correlated. LDA was first proposed as a graphical model for topic discovery and described as a generative probabilistic model that allows sets of observations to be explained by unobserved groups which explain the reasons of similarity in regards to multiple paired components of the data.

3.1 Latent Dirichlet Allocation (LDA)

There are several assumptions behind the LDA Topic Model. (a) Typically not many, but multiple topics exist in the documents. (b) LDA is a probabilistic model with a corresponding generative process which assumes that this simple process demonstrates within each document. (c) A topic is a distribution over a fixed vocabulary like we mentioned in the previous section. Topics are assumed to be generated before the documents. (d) Only the number of topics can be specified ahead.

The generative process of LDA topic model can be written more formally with some notations

- For topics, $\beta_{1:K}$ are from where each β_k is a distribution over the vocabulary
- For document d, θ_d are the topic proportions
- For topic k in document d, $\theta_{d,k}$ is the topic proportion
- For document d, z_d are the topic assignments

- For word n in document d, $z_{d,n}$ is the topic assignment
- For document d, w_d are the observed words

The joint distribution of the hidden and observed variables can be written as

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D} = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \tag{3.1}$$

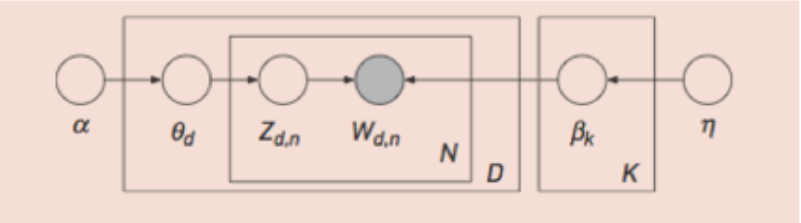

Fig. 1. Plate diagram of the LDA model[22]

There are several objectives we need to know from the plate diagram. (a) On the graph above, only the words which are shaded on the plate diagram are observed. (b) α and η are parameters of the respective dirichlet distributions. (c) Through the process, the topics are generated. (d) plates indicate reiteration of the model [22].

The LDA models are examples of a Bayesian mixture model. In order to understand the estimation of a LDA model with Gibbs sampling for fitting, the model used in this paper, we need to know what the Bayesian model is created with. This explanation is intended to remind and extend your knowledge of probability. The easiest way to understand the Bayesian model is to comprehend Bayes' theorem.

Bayes' theorem is related to conditional probabilities and a formula that describes how to update the probabilities of hypotheses when given evidence. Conditional probability is the probability of one event being true given that another event is true. It is distinct from joint probability, which is the probability that both events are true without knowing that one of them must be true.

Bayes' theorem is connected to our topic modeling. The LDA uses Bayes' formula to find out the relations between words and topics. There are various algorithms for topic modeling we can use in the R software along with **topicmodels**, **quanteda**, **stm**, and **lda** package; Gibbs sampling and variational expectation-maximization (VEM). Gibbs sampling, or a Gibbs sampler, is a Markov Chain Monte Carlo (MCMC) algorithm for obtaining sequences of observations which are approximated from a specified multivarate probability distribution when direct sampling is difficult. As a technical aspect, the R package **lda** (Chang 2011[23]) provides collapsed Gibbs sampling methods for LDA and related topic model variants, with the Gibbs sampler implemented in C. In order to determine the posterior probability of the latent variables all of the models in the **lda** package are fitted using Gibbs sampling.

LDA is identical to probabilistic latent semantic analysis in that each document may be viewed as a mixture of various topics. However, the topic distribution is assumed to be a sparse Dirichlet prior in LSA. The sparse Dirichlet priors encode the data to provide insight such as where documents

cover a small set of topics, or the topics use only a small set of words frequently. In practice, it shows a more precise assignment of documents to topics and less ambiguity with words.

In addition, Semi-collapsed VEM is used for estimation of the Structural Topic Model (STM) in the stm package. It takes sparse representation of a document-term matrix, a non-negative number of topics and covariates, and restores fitted model parameters [24]. Key innovation of Structural Topic Model is to allow users to incorporate arbitrary metadata, which is defined as information about the document in the topic model. Structural Topic Model permits researchers to discover topics and estimate their relationship to document metadata. Hypothesis testing of these relationships can be conducted with this algorithm.

3.1.1 Structural Topic Model

The generative model begins at the top with document $d \in 1...D$ and the words within the documents by $n \in 1...N_d$. Words $w_{d,n}$ from the observations that are instances of unique terms from a vocabulary of terms, which is indexed by $v \in 1...V$. Additionally, the number of topics K is indexed by $k \in 1...K$. Two design matrices provide additional observed information; topic prevalence and topical content, which the analyst can specify given a document which includes a vector of covariates in each row. The matrix of topic prevalence covariates is denoted by $\mathbf{X}$, and has dimension $D \times P$. The matrix of topical content covariates is denoted by $\mathbf{Y}$ and its dimension is $D \times A$. Rows of these two matrices can be denoted by $\mathbf{x}_d$ and $\mathbf{y}_d$. Lastly, marginal log frequency of term v in the vocabulary which can be estimated from total counts defines m_v [25].

The generative process for each document for a STM model with k topics with scalar hyper-parameter s, r, ρ, and K-dimensional hyper-parameter vector σ can be summarized as:

$$\boldsymbol{\gamma}_k \sim \text{Normal}_P(0, \sigma_k^2 I_P), \text{ for } k = 1...K-1, \tag{3.2}$$

$$\boldsymbol{\theta}_d \sim \text{LogisticNormal}_{K-1}(\Gamma' \mathbf{x}'_d, \Sigma), \tag{3.3}$$

$$\mathbf{z}_{d,n} \sim \text{Multinomial}_K(\theta_d), \text{ for } n = 1...N_d, \tag{3.4}$$

$$\mathbf{w}_{d,n} \sim \text{Multinomial}_V(\mathbf{B}\mathbf{z}_{d,n}), \text{for} n = 1...N_d, \tag{3.5}$$

$$\beta_{d,k,v} = \frac{\exp(m_v + k^{(t)}_{k,v} + k^{(c)}_{y_d,v} + k^{(i)}_{y_d,k,v})}{\Sigma_v \exp(m_v + k^{(t)}_{k,v} + k^{(c)}_{y_d,v} + k^{(i)}_{y_d,k,v})}, \text{ for } v = 1...V \text{and} k = 1...K, \tag{3.6}$$

where $\Gamma = |\gamma_1| \cdots |\gamma_K|$ is a $P \times (K-1)$ matrix of coefficients for the topic prevalence model is specified by 3.2 and 3.3. 3.6 is related to $k^{(t)}_{\cdot,\cdot}, k^{(c)}_{\cdot,\cdot}, k^{(i)}_{\cdot,\cdot}$ which is a set of coefficients for the topical content model. Equations 3.4 and 3.5 link to the core language model.

By using the logistic normal distribution, the core language model is able to create correlations in the topic proportions [26][27]. For a model with K topics, we can depict $\eta_d \sim \text{Normal}_{K-1}(\mu_d, \Sigma)$ for the Logistic Normal and map to the simplex with $\theta_{d,k} = exp(\eta_{d,k})/(\Sigma_{i=1}^{K} exp(\eta_{d,i}))$ where $\eta_{d,k}$ is fixed to zero in order to explain the model traceability. For each word within a document d, a topic is sampled from a multinomial distribution $z_{d,n} \sim$ multinomial(θ_d) with given topic proportion vector, θ_d. As a word is chosen from the appropriate distribution over terms B$z_{d,n}$, conditional on the topic can be written $\beta_{zd,n}$ for ease of notation. In the Correlated Topic Model [28], all documents share global parameters both μ and B, but a function of document-level covariates are designated in the structural topic model (STM) [29].

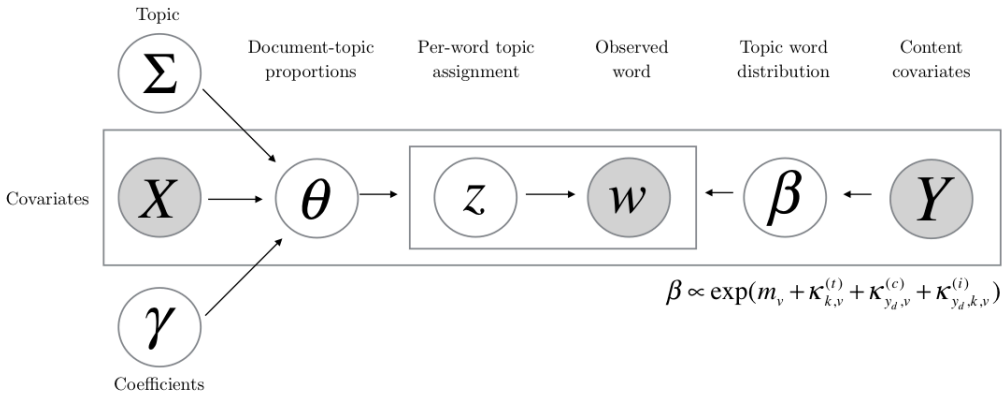

Fig. 2. Illustration of the Structural topic model [29]

3.1.2 Model initialization and model selection for a fixed number of topics

Since the posterior is incurable and non-convex in all mixed-membership topic models, a multimodal estimation is caused that is sensitive to initialization. Starting values of the parameters, such as the distribution over words for a particular topic, affect the results through the estimation procedure [30]. There are two initialization types to deal with this in the **stm** package. One is to choose an initialization based on the method of moments under reasonable conditions for consistency [31]. Since it applies a spectral decomposition (non-negative matrix factorization) of the word cooccurrence matrix, it is known as a spectral initialization. In the **stm** function, it can be chosen by setting **init.type = "Spectral"**. This function will temporarily subset the vocabulary size for the initialization period, if the vocabulary is greater than 10,000 words. Secondly, a short run of a collapsed Gibbs sampler for LDA can initialize the model. It provides a balance of speed, reproducibility, and quality with the **ini.type= "LDA"** in the **stm** function. Because of consistency of producing the best result, generally, the spectral initialization is recommended [30].

If the user or researcher cannot use the spectral initialization, they should estimate several models, randomly generated from starting values, and then evaluate each model through some separate standard. The function, **selectModel**, automates this process to assist the progress of finding a model with desirable properties. Users can specify the number of "runs". **selectModel** creates a net where "run" models are run for two EM steps and the number of "runs" are below 10. Then models with low likelihoods are discarded. The default setting returns the 20% of models with the highest likelihoods, which are run until convergence or reaching the maximum of EM iterations [30]. The FREX metric is developed to measure exclusivity in a way that balances word frequency. FREX is the weighted harmonic mean of the word's rank in terms of exclusivity and frequency under a topic with the empirical CDF exclusivity to that topic. Denoting the KxV matrix of topic-conditional term probabilities as $\mathbf{B}$, the FREX statistic is defined as.

$$\text{FREX}_{k,v} = \left(\frac{\omega}{\text{ECDF}(\frac{\beta_{k,v}}{\Sigma_{j=1}^{K} \beta_{j,v}})} + \frac{1-\omega}{\text{ECDF}(\beta_{k,v})} \right)^{-1} \tag{3.7}$$

where ECDF is the empirical CDF and ω is the weight which is a default value .7 to favor exclusivity in the function [31]. This criteria is calculated for each topic within a model run. There is a plugin estimator for the FREX statistics using the collection **B** coefficients estimated using variational EM [29].

STM allows users to insert a specified number of topics as given function **searchK**. Even though there is no proper answer to define the number of topics that are fitting to a given corpus, the function **searchK** uses a data-driven reach selecting the number of topics [32]. Several automated stats are performed to choose the number of topics including calculating the held out likelihood [33] and executing a residual analysis [34]. For instance, with stable spectral initialization, one could classify a STM model for 7 and 10 topics and compare the resulting along each of the criteria. this **searchK** can also calculate a range of quantities of interest, including the average exclusivity and semantic coherence [30]. In this paper, we apply 6 topics as we introduced in the previous section, but depending on the results such as topic correlations and categories, we may reduce the number of topics to compare results.

4 The Community Views on Criminal Justice System Initiative

The Gun Involved Violence Elimination (GIVE) initiative is managed by the New York State Division of Criminal Justice Services. GIVE uses a variety of evidence-based violence reduction strategies in 20 jurisdictions across New York State. This community views on the criminal justice system project has two objectives. The first is to collect the community's perceptions and experiences with the police department and criminal justice system within one of the GIVE jurisdictions. The second is to share findings with the local GIVE jurisdiction and provide feedback to the community. The research team formulated six main topical categories for which they sought to discover the views of the criminal justice system in the local jurisdiction. Over the course of one year, nineteen focus groups were created across multiple backgrounds founded in police-citizen groups, reentry groups, and community, neighborhood, and youth organizations. These groups had discussion sessions across six major topics and subsequently answered survey questions. Taking into account the potential biases of the note takers and facilitators, and only referencing the quantitative results of the questions, the research team attempted to identify how people answered within those six topics [35].

4.1 Major Topical categories

The six major topics were defined as (i) Interactions with the local police (i) Overall safety (iii) Community concerns (iv) Trust and Fairness (v) Dignity and Respect (vi) Body-worn cameras. With these six topics as their guiding hypothesis, the research team designed qualitative research with procedural justice concepts informed questions. The questions cover police officers and the criminal justice system in a macrocosm. There were three steps of collecting the data in the focused group meetings: rate the questions or statements, display the result of the polls, then sharing interviewees' experiences and opinions along with questions or statements. These following questions and statements are used for the meeting: "Would you describe your most recent interaction with the police as good, bad, neither good nor bad, or no interaction? Who started the interaction? How safe do you feel in your neighborhood at night? What influences this? Overall, how satisfied are you with police responses to community concerns? (Responses range from very satisfied to very unsatisfied.) Rate how strongly you agree or disagree with these statements: I trust the police to do what is best for the community. Overall, the criminal justice system (police, courts, probation, prisons, parole, etc.) tries to do what is best for the community. The police in my community generally treat

people with dignity and respect. The criminal justice system generally treats people fairly. The use of body-worn cameras is good for the relationship between police and this community" [35]. Text mining is an attractive alternative to the traditional constrained Likert-type questionnaires. It is hoped that by letting the community respondents freely expose their views, we can, using modern sophisticated text mining techniques, extract the topic underlying authentic perception of policing and the criminal justice system in general.

5 Statistical Data Analysis

5.1 Sentiment analysis

Sentiment Analysis, also known as opinion mining, refers to the use of natural language processing. As research interest in natural language processing grows up rapidly, it seeks to better understanding on sentiment or opinion expressed in text. The discovery of opinions reflecting people's attitudes upon various topics allows many effective applications, and it is another motivation of sentiment analysis. Semantic orientation (SO) is to measure subjectivity and opinion in text data. It usually catches positive or negative factors and the degree to which the word, phrase, sentence, or document in the data is positive or negative towards a subject topic or idea. Furthermore, text can perhaps be characterized by some other more nuanced emotion like surprise or disgust [36].

Much of Criminal Justice research contains human interactions with the criminal justice system. The data often include people's point of views, expectation, experiences, and concerns towards criminal justice system. The data collected for these types of research include the emotional intent of words. Therefore, we can infer whether a section of text is positive or negative, or characterize more nuanced emotion like surprise or disgust through the sentiment analysis.

In R software, there are three general-purpose lexicons we can use: AFINN, BING, and NRC. All lexicons use unigrams such as single words. These lexicons contain many English words, and words are selected into different sentiment groups or categories. AFINN is a product created by Finn Arup Nielsen. This lexicon assigns words with a score that fits between -5 and 5, which matches negative scores implying negative sentiment and positive scores implying positive sentiment. The BING lexicon was created by Bing Liu and collaborators and indicates polarity of the words in a sentence. It categorizes words into negative and positive categories. Lastly, the NRC lexicon was created by Saif Mohammad and Peter Turney. NRC lexicon assigns words into binary values ("yes"/"no") fitting into eight categories: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust [37].

5.1.1 Data exploration with word cloud

A tag cloud (word cloud, or weighted list in visual design) is a visual representation of text data, typically used to describe keyword metadata (tags) on websites or to visualize free form text [38]. A word cloud or text cloud is a visualization of word frequency in a given text as a weighted list [39]. This technique has been popularly used to visualize the topical content of political speeches and several journal papers to visualize their text data [40][41][42]. The world cloud for this paper is made in R version 3.2.2 using packages 'SnowballC', 'wordcloud', and 'RColorBrewer'. In wordclouds, the size and color of each word is determined by the frequency of its appearance in the a list, in this case, the major 250 words were sorted by its frequencies to depict over all documents. The words which appeared most frequently were placed closer to the center of the document and displayed using a larger font size. The words were also sorted based on coloration such that from most frequent to least frequently occurring, the words were colored black, yellow, pink, blue, orange, and green. We can display words randomly as well. The rotation of words and display options also can be changed.

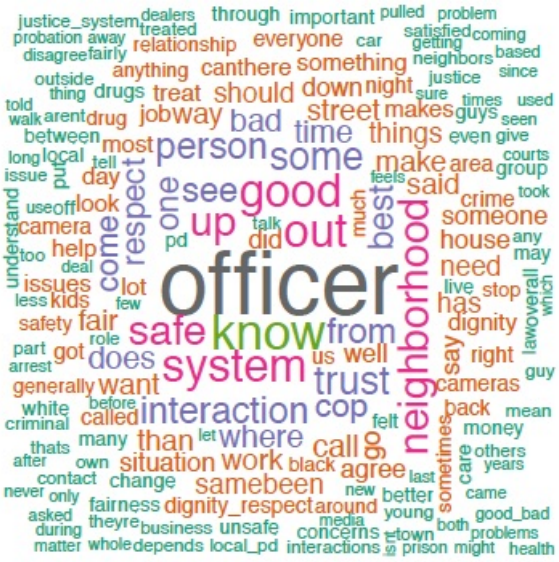

Fig. 3. Word cloud displayed by frequencies of text

Fig. 4. Word cloud sorted by sentiment categories; disgust, fear, anger, joy, trust, and sadness

5.1.2 Wordcloud from sentiment snalysis

In this paper, we applied all of lexicon techniques to our sentiment analysis. The above word cloud was created with only six categories to present what words appear in the matrix. The difference between the previous word cloud and this word cloud is the dependence on the frequency of each word in the matrix. Sentiment clouds do not sort the words by frequency. Colors are used to display categories. The font sizes differ by categories and it is an unsupervised technique, meaning that they are automatically generated. Therefore, the font size should not be considered when

interpreting the sentiment cloud. With this word cloud, we use the exact words our interviewees used when they described their own experiences. The reason we chose the above six categories was that these are the most common positive and negative emotional categories.

Fig. 5. Word cloud sorted by polarity sentiment categories; positive and negative

The above word cloud was created by the BING lexicon and its dictionary. Words were sorted by negative and positive categories in the BING dictionary and displayed in the cloud.

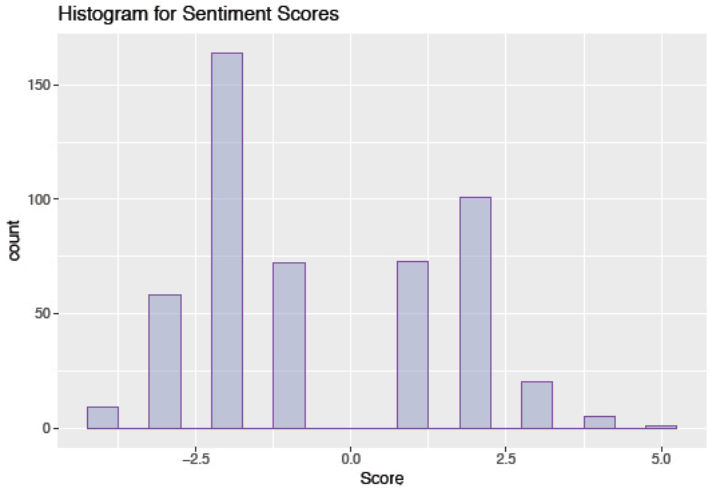

Fig. 6. Sentiment score through Afinn

With bigrams, Afinn dictionary scored each sentiment word between -5 to 5. Through the Fig. 8, we can see more negative scores than positive words. There is no -5 scored words which indicates extremely negative words, but most words are light positive like between 1 and 2 or scored mostly negatively. 60% of the words are negatively scored, no neutral words such as scored zero, and only 39.76% which is almost 40% are positive. Summarized score is -270 even though the average of the negative score is 2 and average of the positive score is 1.8 which means there is not much difference between scores, but there are much more negative words than positive words in the data set.

5.2 Cluster of documents

Cluster analysis categories data into groups (clusters) that are meaningful, or useful, or both. The clusters should capture the natural structure of the data, if meaningful groups are the goal. However, only data summarization is useful. Also, cluster analysis is a multivariate method which targets to classify sample objects from the base of a set of measured variables into similar subjects are placed in the same group [43]. Cluster analysis allows one to obtain an abstraction from individual measured data objects (or subjects) to the clusters of which those data objects consist. In addition, some techniques classify each cluster in terms of a cluster prototype. For instance, a data object that represents the other objects in the cluster [44].

There are a number of different applications that we can use for cluster analysis; these applications can be described as the following two categories:

- Hierarchical methods
  - Agglomerative methods which objects start in their own independent cluster. Then the most similar clusters are combined and this process is repeated until all objects belong into one cluster. Eventually, the optimum number of cluster is chosen from all cluster solutions.
  - Divisive methods which all objects start in the same cluster and agglomerative strategy is implied reversely until every object is in an independent cluster. Agglomerative methods are more popular than divisive methods.
- Non-hierarchical methods (known as k-mean clustering methods)

5.2.1 Hierarchical cluster analysis

For our hierarchical clustering analysis, we adapted the euclidean distance to measure the distance and ward's method among agglomerative methods to determine which clusters should be joined at each stage. The most common illustration tool for this hierarchical cluster analysis is a dendrogram. This diagram illustrates which clusters have joined at each stage of the analysis and the distance among clusters when it joins [43].

In the above dendrogram, we set to demonstrate four numbers of groups (clusters). It seems like we can reduce number of clusters and the distances are too small. Therefore, we decided to use exploratory graph analysis via the R-package EGA. It is currently under development, but it is quite effective to show networks between clusters. It shows directly what items or documents are belong to what clusters.

The dendrogram of hierarchical cluster analysis and EGA are intuitively similar, but EGA solution seems to be more reasonable. We would like to know what types of disparate factors they have as two major groups through sentiment analysis. However, there were no differences between the two groups in sentiment analysis. Both were majorly negative.

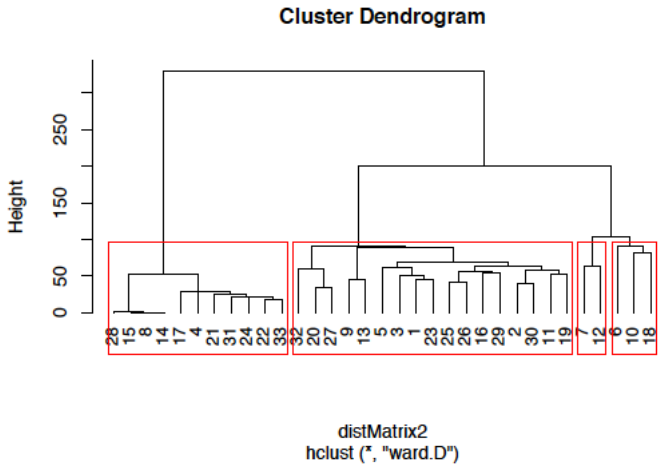

Fig. 7. Dendrogram of document clusters

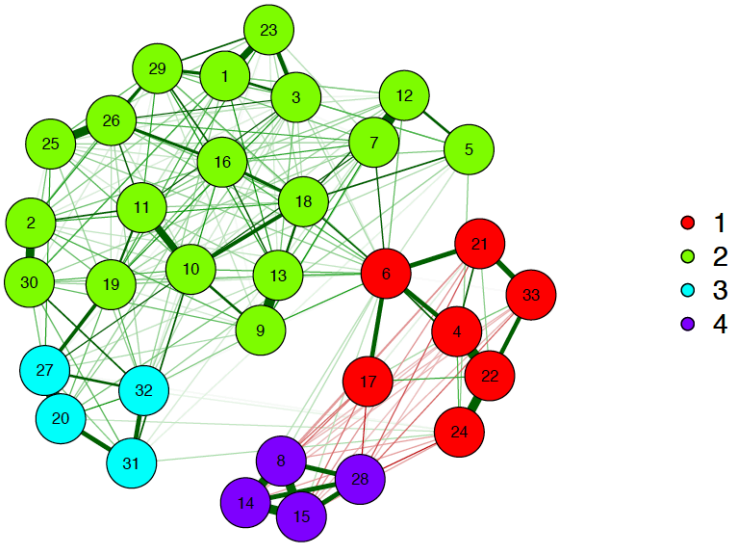

Fig. 8. Exploratory Graph Analysis (EGA)

5.3 Topic modeling with n-gram via Structural Topic Model (STM)

The most frequent terms within a topic summarize the majority of topic models, despite that there are several methods for choosing higher order phrases [45][46]. We use a metric referred to as FREX to summarize topics that associate frequency term and exclusivity to that topic into a univariate summary statistic [47][48].

Based on results from STM, we classified which topics can fit into research topics. Body-worn camera project in the city police department started after an unfortunate incident so that we

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Cop, gun, care, control, youth, kids, young	Job_well, confidence, conduct, safety, criminal_justice, describe, confidence_confidence	Individuals, told, check, called, noise, women, shut	Understand, black, which culture, media, school, trained	Otpolice, sees, tlocal, tlocal_pd, illness, feels, thinks	Health_center, guys_street, fired, polite, guys, business, dealers

Fig. 9. Six topics through STM

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
Interaction	V					
Safety		V				
Community concerns			V		V	V
Trust and fairness						
Dignity and respect				V		
Body-worn camera						

Fig. 10. Fitting in original research topics

expected to catch community views on the project. Unfortunately, there is no direct evidences that many citizens in the community are interested in Body-worn cameras. Most of topics belonged to community concerns. Others are interaction, safety, and dignity and respect. Therefore, we decided to reduce the number of topics. The following table shows what can be affected, if we reduce the number of topics, such as fitting the topics. Additionally, through the results, we assume that several topics are related to each other such as interaction, trust and fairness, and dignity and respect.

Topic 1	Topic 2	Topic 3	Topic 4
Cop, sense, control, care, youth, gun, training	Confidence, conduct, contact_interaction, neighborhood_problems, car_street	House, drugs, told, individuals, residents, especially, fired	Health_center, guys_street, clinic, guys, business, business_owners, owners

Fig. 11. Four topics through STM

Even though we reduced the number of topics, the contents did not change. Also, categories where the topics belong are the same, which are interaction, safety, community concerns, dignity and respect. We began to wonder as it seems like the most topics are talking about interactions with cops and officers and neighborhoods and concerns inside of them. Therefore, we would like to reduce once more into two topics.

The Fig. 12 represents when we narrow the number of topics down into two, we see one topic is related to criminal justice system like court, cops, and officers whereas, the other topic shows how they get along with the community and neighborhood.

Topic 1	Topic 2
Neither, neither_good, social_media, moving, reputation, base, brought	Confidence, affects, overall_satisfied, released, courts_probation, conduct, contact_interaction

Fig. 12. Two topics through STM

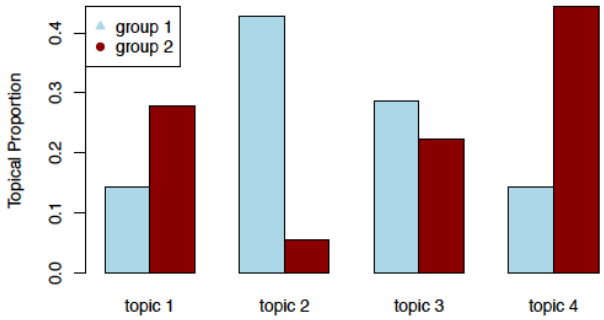

Fig. 13. Difference via Topics between Two Major Clustering

In the previous section, we applied clustering analysis on our data set. Fig. 10, the EGA graph, presents two major clusters (groups): group 1 and 2 in the graph. We would like to detect the difference between the two groups with topic model. The figure 15 presents topical proportions via four topics between two groups. There is the obvious difference between two groups. Especially, the two groups which show total opposites in the topic 2 and 4.

6 Discussion

In the section 5.1, we could discover what types of words with which community residents describe their experiences with the criminal justice system in their daily life or what they heard from their neighbors, friends, and family members. People used more negative words than positive words, like 60% of negative and zero neutral sentiment were captured based on the figure 8. Also, through clustering analysis in the section 5.2, we could demonstrate four clusters (or groups), but it can be concluded as two major groups. Sentiment analysis could not make distinct two major groups since it had the exact same patterns of the sentiment in NRC, but figure 15 in the section 5.3 shows distinction between the two major groups. The two groups have absolute opposite topical proportions in the each topic. Furthermore, section 5.3 STM shows there are four major topics in the data set. Also, we can narrow down into two topics based on the results, but as a research topical perspective, four topics seem to be the best since there is not enough evidence for body-worn cameras and trust and fairness topics from figure 11 to 14 in section 5.3.

The data set has a lot of questionaries and most of text documents do not have format, since there were several research assistants who took the notes during the interview. Each of them have different note taking styles and personal abbreviations of words. It means that the data sets have a lot of noise and it required many steps of preprocessing. Therefore, we need to format the note taking for the future. The formatting does not include labeling of the documents, including name of the group, name of note takers, and location of the interview. Also, we need more data to approach and conclude what the community really thinks about the criminal justice system. With our current

data set, which is 33 valid text documents, there are only a few things we can conclude. Even if we applied as many applications as analysis data, if the data set is small, we cannot conclude or extract the right information as what the data set would contain.

None of the topics include body-worn cameras. It presents two things. The first one is body-worn cameras is not the dominant topic in the data set. For instance, some interviewees mentioned about body-worn cameras during the interview, but it was not much. The second is that people do not know, or have no idea, about body-worn cameras. The research team asked the same questions to all interview groups so that all document has questionaries related to body-worn cameras. If the topic modeling analysis cannot catch this topic, it may means that interviewees do not have any opinion about body-worn cameras.

In this paper, we applied n-gram with Structural Topic Model (STM), but Hidden Markov Topic Model (HMTM) can be more effective with this social science data sets. HMTM goes beyond the common bag-of-words paradigm, and infers semantic representations by taking into account the hereditary sequential nature of linguistic data [49]. Also, it is an enhanced Bayesian bag-of-words model so that it can catch topics from text data sets that have a complexity problem.

7 Conclusion

In our research, we integrated text mining techniques into the text document data from the CVCJS initiative, such as n-gram, sentiment analysis, and structural topic model (STM) along with a multivariate analysis technique, clustering analysis. Through our statistical approach, we aim to identify the community views on the criminal justice system in one of the 20 New York State GIVE jurisdictions. We could conclude a couple of things, such as the number of topics we could discover through analysis, which was less than the actual research target topics, overall sentiment of neighborhood, and number of clusters (or groups) exists and its similarity and disparity. However, there are several inconclusive and limited results due to the data format, small numbers of text data, and other factors.

Acknowledgement

This research was supported by the Center for Public Safety Initiative (CPSI) at the Rochester Institute of Technology (RIT) Department of Criminal Justice. Of the research team at CPSI, we would like to acknowledge Dr.John Klofas, who is the director of CPSI and the principle investigator of the CVCJS initiative, as well as their research associate, Mary Beth Spinelli, and research assistants Christina Burnett, Chaquan Smith, Avanelle Bernard, and Jamie Dougherty.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Samuel Walker. Origins of the contemporary criminal justice paradigm: The American bar foundation survey. 1953-1969. Justice Quarterly. 1992;9(1):47-76.

[2] President's Commission on Law Enforcement and Administration of Justice. The challenge of crime in a free society. US Goverment Printing Office; 1967.

[3] Chapman R, Scheider M. Community policing defined. U.S. Department of Justice-COPS; 2012.

[4] Yili Xu, Mora L Fiedler, Karl H Flaming. Discovering the impact of community policing: The broken windows thesis, collective efficacy, and citizens' judgment. Journal of Research in Crime and Delinquency. 2005;42(2):147-186.

[5] Laura Merkey. Building trust and breaking down the wall: The use of restorative justice to repair police-community relationships. Mo. L. Rev. 2015;80:1133.

[6] Salim Alami, Omar Elbeqqali. Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts. In Intelligent Systems: Theories and Applications (SITA), 2015 10th International Conference on, pages 1–5. IEEE; 2015.

[7] Matthew S. Gerber. Predicting crime using twitter and kernel density estimation. Decision Support Systems. 2014;61:115-125.

[8] Jacopo Soriano, Timothy Au, David Banks. Text mining in computational advertising. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2013;6(4):273-285.

[9] David Meyer, Kurt Hornik, Ingo Feinerer. Text mining infrastructure in r. Journal of Statistical Software. 2008;25(5):1-54.

[10] Martin Rajman, Martin Vesely. From text to knowledge: Document processing and visualization: A text mining approach. In Text mining and its applications: Results of the NEMIS Launch Conference. Springer Science & Business Media. 2004;138:7-24.

[11] Jiliang Tang, Xufei Wang, Huiji Gao, Xia Hu, Huan Liu. Enriching short text representation in microblog for clustering. Frontiers of Computer Science in China. 2012;6(1):88-101.

[12] Panagiotis Kanellis. Digital crime and forensic science in cyberspace. IGI Global; 2006.

[13] Cooper K. Dolphins, aliens, and the search for intelligent life, astrobiology magazine; 2011.

[14] Vladimir Vapnik, Akshay Vashist. A new learning paradigm: Learning using privileged information. Neural networks. 2009;22(5):544-557.

[15] David M Blei, Andrew Y Ng, Michael I Jordan. Latent dirichlet allocation. Advances in Neural Information Processing Systems. 2002;601-608.

[16] David Banks. Mining text networks. UP-STAT 2015: Fourth Joint Conference of the Upstate Chapters of the American Statistical Association; 2015.

[17] Nigam Kamal, McCallum Andrew Kachites, Thrun Sebastian, Mitchell Tom. Text classification from labeled and unlabeled documents using EM. Machine Learning. 2000;39(2):103-134.

[18] Deerwester Scott, Dumais Susan T, Furnas George W, Landauer Thomas K, Harshman Richard. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41(6):391.

[19] Broder Andrei Z, Glassman Steven C, Manasse Mark S, Zweig Geoffrey. Syntactic clustering of the web. Computer Networks and ISDN Systems. 1997;29(8-13):1157-1166.

[20] Hofmann Thomas. Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 1999;50-57.

[21] Jordan MI Blei DM, Ng AY. Latent dirichlet allocation. Journal of Machine Learning Research. 2003b;3:993-1022.

[22] Stephen Clark. Topic modelling and latent dirichlet allocation. Online, Lent; 2013.

[23] Chang Jonathan. lda: Collapsed Gibbs sampling methods for topic models. Online; 2011.

[24] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, David G Rand. Structural topic models for open-ended survey responses. American Journal of Political Science. 2014b;58(4):1064-1082.

[25] Edoardo M Airoldi, William W Cohen, Stephen E Fienberg. Bayesian models for frequent terms in text. In Proceedings of the Classification Society of North America and INTERFACE Annual Meetings. 2005;990:991.

[26] John Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological). 1982;139-177.

[27] Atchison J, Sheng M Shen. Logistic-normal distributions: Some properties and uses. Biometrika. 1980;67(2):261-272.

[28] David M Blei, John D Lafferty. A correlated topic model of science. The Annals of Applied Statistics. 2007;17-35.

[29] Margaret E Roberts, Brandon M Stewart, Edoardo M Airoldi. A model of text for experimentation in the social sciences. Journal of the American Statistical Association. 2016b;111(515):988-1003.

[30] Margaret E Roberts, Brandon M Stewart, Dustin Tingley. Stm: R package for structural topic models. (Working paper).

[31] Margaret E Roberts, Brandon M Stewart, Dustin Tingley. Navigating the local modes of big data. Computational Social Science. 2016a;51.

[32] Justin Grimmer, Brandon M Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political Analysis. 2013;21(3):267-297.

[33] Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, David Mimno. Evaluation methods for topic models. ACM. 2009;1105-1112.

[34] Matt Taddy. On estimation and selection for topic models. In International Conference on Artificial Intelligence and Statistics. 2012;1184-1193.

[35] Spinelli MB, Smith C, Klofas J. Measuring community views on the criminal justice system with group feedback analysis. Technical report, Rochester Institute of Technology; 2017.

[36] Charles Egerton Osgood, George J Suci, Percy H Tannenbaum. The Measurement of Meaning. University of Illinois Press; 1964.

[37] Julia Silge, David Robinson. Text Mining with R: A Tidy Approach. O'Reilly Media, Inc; 2017.

[38] Martin J Halvey, Mark T Keane. An assessment of tag presentation techniques. In Proceedings of the 16th International Conference on World Wide Web. ACM. 2007;1313-1314.

[39] Joe Lamantia. Text clouds: A new form of tag cloud. 2007;10(09):2010. Available:http://www. joelamantia. com/blog/archives/tag_clouds/text_clouds_a_new_fo rm_of _tag_cloud. html¿ Acesso em

[40] Chirag Mehta. Us presidential speeches tag cloud; 2007.

[41] Joseph B Sempa, Eva L Ujeneza, Martin Nieuwoudt. Systematic review of statistically-derived models of immunological response in HIV-infected adults on antiretroviral therapy in Sub-Saharan Africa. Public Library of Science. 2017;12(2):e0171658.

[42] Soo Downe, Kenneth Finlayson, Tunçalp, A Metin Gülmezoglu. What matters to women: A systematic scoping review to identify the processes and outcomes of antenatal care provision that are important to healthy pregnant women. BJOG: An International Journal of Obstetrics & Gynaecology. 2016;123(4):529-539.

[43] Everitt BS, Landau S, Leese M. Cluster analysis. Hodder Arnold Publication. Wiley; 2001.

[44] Tan PN, Steinbach M, Kumar V. Introduction to data mining. Pearson Education, Limited; 2014.

[45] Qiaozhu Mei, Xuehua Shen, ChengXiang Zhai. Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. 2007;490-499.

[46] David M Blei, John D Lafferty. Visualizing topics with multi-word expressions. 2009; arXiv preprint arXiv:0907.1013.

[47] Allison June-Barlow Chaney, David M Blei. Visualizing topic models. In ICWSM; 2012.

[48] Airoldi EM, Bischof JM. A regularization scheme on word occurrence rates that improves estimation and interpretation of topical content (with discussion). Journal of American Statistical Association; 2016.

[49] Mark Andrews, Gabriella Vigliocco. The hidden markov topic model: A probabilistic model of semantic representation. Topics in Cognitive Science. 2010;2(1):101-113.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/22537