Hien Nguyen and Shraeya Srinivasan

Sentiment Analysis using LDA on Random Text Entries to Network Visualization

23rd April 2022

Abstract

Sentiment recognition and analysis is a growing area in multi-disciplinary research that has become increasingly popular in recent years. It is especially valuable in assessing public attitudes such as in customer service areas, in monitoring brand and product sentiments from customer feedback, and can also be used to assess sentiments surrounding different marketing campaigns. Especially with a constant phenomenon of information overload online, the role of sentiment analysis models is more crucial to provide a quick overview of public opinions. This paper discusses application of LDA topic modeling, particularly HDP-LDA, and explores Naive Bayes methods for emotion analysis and the preprocessing needed for these models. The goal of the exercise is to visualize a network graph by separating nodes into emotions and sub-nodes into topics of different interests. The experiment was proven to be successful as both models were able to predict a various number of topics that could be sorted across many types of emotions. This paper can be considered as a foundational step for further experimental refinement to get more accurate results in sorting random texts into emotions and topics.

Introduction and Motivation

Topic modeling is a useful tool to explore the underlying trends in textual data, and in this case to understand underlying sentiment and emotions. Latent Dirichlet Allocation (LDA) is a generative probabilistic statistical approach to document modeling that takes a set of texts and assesses each topic as a mixture over a vocabulary set of words and each overarching document as a mixture over topic probabilities. It is an unsupervised method that can identify structure of key topic word groups in text data and the number of topics as some output parameter. Rather than grouping the documents, the model groups the words within the texts.
Hierarchical Dirichlet Processes (HDP) is a Bayesian nonparametric (BNP) method in which the parameters, or in this case the number of topics, can be inferred from the data rather than pre-define the number of topics. In Dirichlet process mixtures, the topics are learned from the data and HDPs allow for an infinite number of topics that are only bounded by the vocabulary of the data and can be used across the documents.

Naive-Bayes (NB) models utilize the Bayes theorem of probability to classify a dataset and assign probabilities for each class based on the likelihood that a given record belongs to that class. It assumes independence between the different features, or predictors.
Sentiment analysis models can study public attitudes towards some entity or text and can generally be divided into emotions that are positive, negative, and neutral forms. Emotions and mood can comprise a range of feelings, behaviors, outlooks, and mental states and can often be very subjective. This project drew inspiration from the inner layer of six main emotions, outlined (Appendix A, **Fig. 1**) and focuses on anger, fear, joy, love, sadness, and surprise for the LDA and NB Soft and Hard Expectation-Maximization classification methods. The following sections discuss the experimental setup for the LDA, HDP-LDA, and NB models, along with the network graph visualization, output results, and conclusions and future exploration with this research topic.

Experimental Setup

This project aimed to visualize a network of analyzed texts sorted into emotions and topics found in those texts. First, the data was completely unsupervised. Small paragraphs were then generated randomly using an online generator, and these were populated into an excel sheet to later convert into a data frame for preprocessing and analysis. Sentences and tweets were also populated into the dataframe in order to maintain a data set of a variety of text formats. Random small paragraphs were generated rather than pulled from an existing text in order to ensure that the models can be applicable to different texts and use cases by representing texts from a variety of sources. Each was represented as a separate entry.
We also were able to find a labeled dataset of sentences and the associated emotions. This was later used in training supervised, and semi-supervised Naive Bayes Expectation-Maximization models to predict emotions.

Preprocessing

We first performed a series of preprocessing steps, especially for the LDA models. This included the removal of special characters, punctuation, and capitalization, and tokenization of the data so that each document text would be split into separate words. Words were updated to remove any accents, maintained as bigram phrases if relevant, and key stopwords that are not informative to emotion were removed. Lemmatization was a fairly time-consuming part of the preprocessing but focuses on part of speech to revert words back to their “dictionary” form for better modeling. Word clouds were used to visualize the common words in the dataset, in which words that were uninformative were added to the stop words list.

LDA

We used the inner layer of emotions graph labels to train our emotions LDA model, therefore K is set to 6. Following preprocessing, the tokenized data set was created into a Dictionary, a mapping between words and their integer ids. This Dictionary was made into a bag or words (BOW) corpus and a TF-IDF corpus to be trained. These corpuses were used to train in their respective BOW and TF-IDF models. Each topic was found by identifying key vocabulary words that were assigned a weight, to produce feature vectors. The pyLDAvis package was used to visualize the topics in an intertopic distance map with the top relevant terms for each topic. This visualization method was also an important way to find unrelated keywords to emotion clusters to add in the stop words list and retrain again. Finally, we use a cluster bar graph to see if there were any distinction of count of predictions by clusters of emotions to see how well the trained model is, along with evaluated f1 and v_measure scores. (Appendix B, **Fig. 3**).

HDP-LDA

We used HDP-LDA which would allow us to not have to set an initial K topics value, as we didn't know the maximum bound for the number of topics in randomly generated texts. First texts go through a simple_preprocess method from the gensim library to remove words that have length less than 3. Bigrams are created from the simple_preprocessed word list, in which they are lemmatized for training. If lemmatized texts have no words, we remove the entire entry. To create a HDP-LDA model, we used the python library tomotopy. There are three possible HDP-LDA models, ONE, PMI and IDF, their characteristics are as follows:

Model	**Characteristics**
ONE	All terms inside texts are weighted equally.
PMI	Pointwise Mutual information term weighting.
IDF	Inverse Document Frequency: Down-weights high frequency terms, upweights low frequency ones.

Whilst initializing the HDP model, we restricted the model to use only words that appeared in at least 10 documents with the min_cf whilst excluding the 10 most frequent words with rm_top argument. Since the texts are randomized and relatively small, we believe that words that appear in more than 1 document are fairly common, along with frequent words in documents. The gamma and alpha hyperparameters symbolize whether documents share many topics while individual documents are of few topics, respectively. We also initialized the number of topics

with initial_k to be 50 since we ended up having 65 random paragraphs, 1200+ sentences and a multitude of tweets imputed.

We chose the Monte Carlo inference method to trade-off accuracy for speed/memory for the exercise, therefore, initializing the model, we set the mcmc_iter to be 2000. Once the HDP-LDA models of each term weights are trained on the unsupervised dataset, a number of topics (topics_id) along with topic labels associated with their respective scores are found. We then use a coherence scoring model from the gensim python library to judge coherence of topics for each of the HDP-LDA models with different term weight settings. We then use the highest coherence scoring model to classify topic_ids on the same unsupervised dataset.

Naive Bayesian Classification Methods

Both Hard Expectation Maximization (EM) and Soft EM were implemented on a semi-supervised dataset. Using the labeled emotions dataset, we implemented a SoftMultinomialNB model using MultinomialNB from the sklearn library. The entries of the labeled dataset were vectorized using the CountVectorizer function, in which it removed common English stopwords. The semi-supervised dataset included data from the labeled set with the unsupervised set that was used in the HDP-LDA model. The experiments of each of the models (soft and hard semi-supervised EM, with supervised EM) were evaluated using F1 and V-measure scores.

Network Graph Visualization

We lastly visualized the results into a network graph. To do this, we created a topics dataframe that had the original text entry, the lemmatized version of that entry, the topic_id associated with the entry found through the HDP-LDA, and the top associated keywords for the topic_id. We then used the Hard EM model (best performing at the time) to predict the emotions of each original text (after being vectorized). Finally, the topics_id and the emotions associated with them were counted and the network graph was visualized.

Results and Discussion

LDA

The intertopic distance map in (Appendix B, **Fig. 3**) identified the 6 distinct topics that we chose for K. The most frequent or relevant words could be analyzed to better understand the topic at hand. In the case of topic 2, words such as “love”, “life”, “important”, “free”, and “enjoy”, suggest that the topic could relate to a more optimistic or positive emotion.

V-measure scores from 0 to 1 indicate how well and how homogenous and complete the clustering partition is with 1 indicating that the data in each cluster are similar and also similar data are assigned to the same cluster. The F1 scores from 0 to 1 indicate the models accuracy and precision on the dataset in which the model classifies each record into the correct cluster, with 1 being the highest and best F1 score. For the purpose of this project, we considered a V-measure score and F1 score of above 0.7 to be a generally acceptable score. The V-measure score of our LDA models were extremely low at about 0.0049 and 0.008, indicating very poor clustering (Appendix B, **Table 1-2**).
This can be further seen when looking at the plot of document counts of each emotion topic in each of the generated clusters using the LDA model (Appendix B, **Fig. 4-5**), there was no evident peak to assign any particular cluster for each emotion. Instead the clusters across each emotion were rather uniform and failed to present an obvious cluster assignment. This confirmed a fairly poor evaluation of our LDA model and we decided to further investigate emotion analysis using NB methods as discussed later in this results section.

HDP-LDA

A coherence metric allows us to objectively evaluate how well the word groupings fit a topic that generally makes sense, by identifying how likely the word groupings assigned by the model exist across the text documents. A good coherence metric typically lies within 0.5-0.65, and all of the ONE, Pointwise Mutual Information (PMI), and Inverse Document Frequency (IDF) models had coherence scores around or above 0.5. The IDF and PMI models yielded a slightly higher coherence score (Appendix C, **Fig. 6**). The coherence scores can give us a general idea of how well our latent topics fit. However, in terms of utilizing the model to determine topics on an unseen document, it may not generalize as well because the unseen document is not as directly part of the trained data. Seeing that the PMI model yielded the greatest coherence score, indicating that it would be the best fit model, another thing to note is that the last iteration of the pmi model found about 63 topics, but when actually classifying it appears that not all discovered topics might be used, but rather 37 unique topics were found in the document entries.

NB

For each of the naive bayes models including semi supervised with soft and hard EM, the plots displaying document counts of each emotion topic within each of the generated clusters presented a clear peak, or main cluster for the given emotion being analyzed. In (Appendix D, **Fig. 7**), anger can be clearly assigned as cluster 1, with fear assigned to cluster 4, joy to cluster 5, love to cluster 2, sadness to cluster 0, and surprise to cluster 3. Similarly, the semi-supervised soft and hard EM models also clearly match the given emotions to separate clusters as seen in (Appendix D, **Fig. 8-11**).

Compared to our LDA model with extremely low V-measure scores, our NB models yielded notably higher V-measure scores. Our initial NB model produced a V-measure score of about 0.597 and an F1 score of about 0.781 (Appendix D, **Table 3**). Our semi-supervised soft and hard EM models resulted in V-measures of about 0.767, 0.570, 0.778, and 0.589 (Appendix D, **Tables 4-7** respectively). Particularly for the training data for both soft and hard EM (Appendix D, **Tables 4,6**), the V-measure scores were above 0.7 indicating good clustering homogeneity and completeness. The F1 scores were 0.888, 0.732, 0.894, and 0.743. All F1 scores were greater than 0.7 and considered acceptable for our purposes.

Network Graph Visualization

The network graph visualization (Appendix A, **Fig. 2**) showed basic connections between the emotions analyzed and the topics_id found in the random texts. There is a lot of room for improvement for this network visualization, for example to show the weights of each node by size - if certain topics with the same emotions were found in more than one texts (our analysis.csv shows this but we didn't have time to work with the library to implement). Also better distinct coloring could be used to dictate emotions and topic_ids separately for more clarity.

Conclusions and Future Work

A significant portion for any topic modeling, as discussed in this paper, revolved around data tidying and preprocessing. In addition, it can be difficult at times to infer the topics at hand by just looking at feature vectors, instead it was beneficial to embed more manual inference visualizations such as analyzing word clouds. When exploring the emotions analysis with LDA, we realized that there was difficulty in some of this implementation, and therefore drove us to explore NB methods as well. HDP-LDA proved to be extremely great at finding unknown topics, making it beneficial to dive quickly into the dataset.

Future exploration and application in sentiment analysis across various entities could open up a strong potential area for predicting the attitudes and opinions of the public and infer behavior patterns and reactions. Continuing to improve hyperparameter tuning can also help for better results. We would also want to aim for even greater V-measure and F1 scores for our models. Based on the emotions dataset we used here, we considered scores greater than 0.7 to be reasonable for our purposes, but future work could include building these models with a larger labeled dataset in order to yield stronger V-measure and F1 scores. However, by nature an emotions analysis project might handle a notable amount of overlap across emotions. For example, joy and love could be quite related along with fear and surprise.

Furthermore, since the HDP-LDA only returns topic_ids, and the possible labels associated with them, a manual labeling process could be used to further achieve topic modeling clarity.

Note: Our final code submission focuses on the HDP-LDA and NB applications for our outputs as the highest performing models, but for supplemental exploration, our extensive code implementation can be found at https://github.com/hiennguyen26/Emotions_Topics_Modeling.git

References

Dakshina, K., and Rajeswari Sridhar. “Lda Based Emotion Recognition from Lyrics.” *Smart Innovation, Systems and Technologies*, 2014, pp. 187–194. doi: 10.1007/978-3-319-07353-8_22.

Debashis, Naskar, et al. “Sentiment Analysis in Social Networks through Topic modeling.” *2016 Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

Ding, Wanying, et al. “A Novel Hybrid HDP-LDA Model for Sentiment Analysis.” *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013. doi: 10.1109/wi-iat.2013.47.

“Emotion · Datasets at Hugging Face.” *Emotion · Datasets at Hugging Face*, https://huggingface.co/datasets/emotion.

Farkhod, Akhmedov, et al. “LDA-Based Topic Modeling Sentiment Analysis Using Topic/Document/Sentence (TDS) Model.” *Applied Sciences*, vol. 11, no. 23, 2021, p. 11091. doi: 10.3390/app112311091.

“Gensim: Topic Modelling for Humans.” *Models.hdpmodel – Hierarchical Dirichlet Process - Gensim*, 22 Dec. 2021, https://radimrehurek.com/gensim/models/hdpmodel.html.

Kapadia, Shashank. “Evaluate Topic Models: Latent Dirichlet Allocation (LDA).” *Medium*, Towards Data Science, 29 Dec. 2020, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0.

Kapadia, Shashank. “Topic Modeling in Python: Latent Dirichlet Allocation (LDA).” *Medium*, Towards Data Science, 29 Dec. 2020, https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0.

Limsettho, Nachai, et al. “Comparing Hierarchical Dirichlet Process with Latent Dirichlet Allocation in Bug Report Multiclass Classification.” *15th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2014. doi: 10.1109/snpd.2014.6888695.

Sroka, Eduardo Coronado. “Don't Be Afraid of Nonparametric Topic Models.” *Medium*, Towards Data Science, 16 May 2020, https://medium.com/towards-data-science/dont-be-afraid-of-nonparametric-topic-models-d259c237a840.

Sroka, Eduardo Coronado. “Don't Be Afraid of Nonparametric Topic Models (Part 2: Python).” *Medium*, Towards Data Science, 24 Sept. 2020, https://towardsdatascience.com/dont-be-afraid-of-nonparametric-topic-models-part-2-python-e5666db347a.

Appendix A Project Visualizations

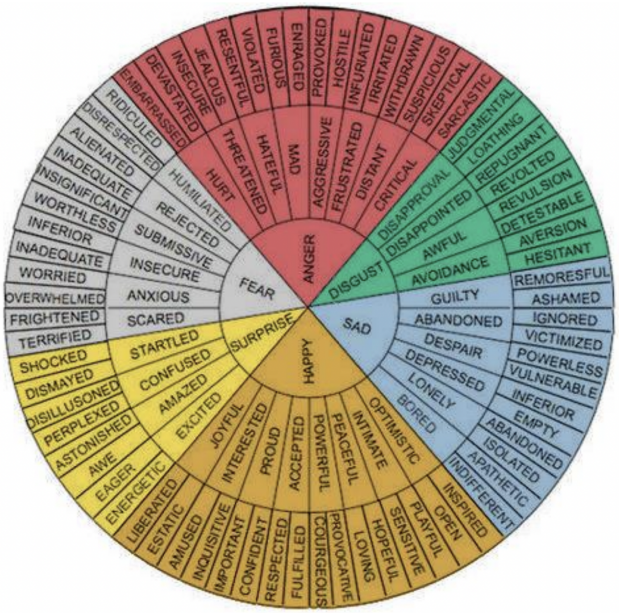


Figure 1: Layers of emotions with 6 main emotions found in the innermost layer

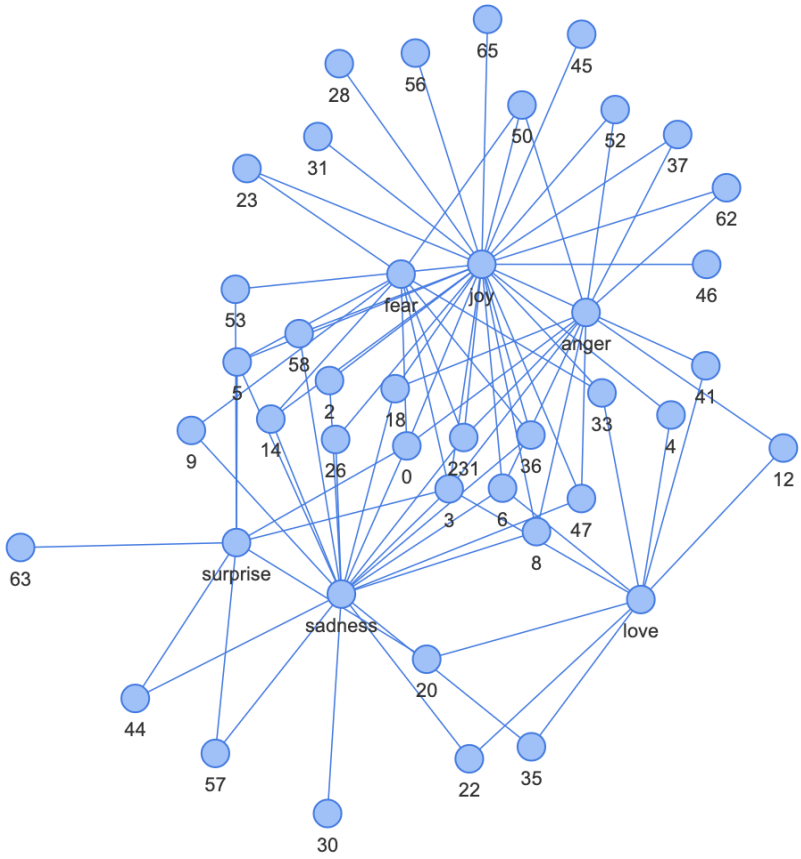


Figure 2: Network Graph Visualization of the counts of topics and the associated emotions

Appendix B Emotions LDA Output Data and Visualizations

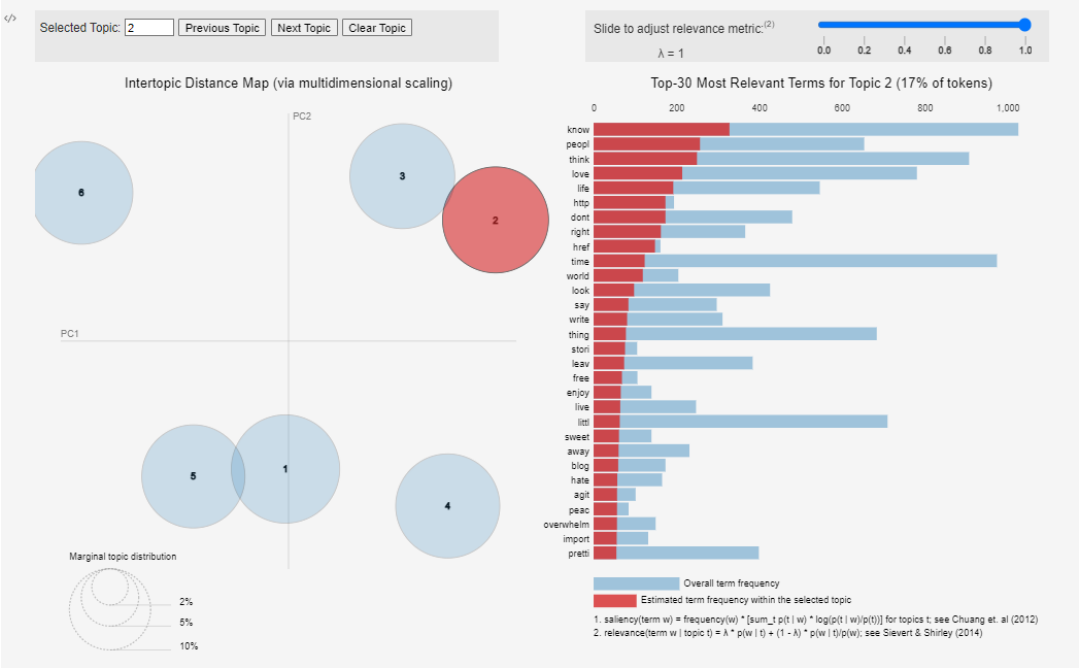


Figure 3: Most frequent/relevant terms by topic with different values for λ parameter, specifically the intertopic distance map and top-30 most relevant terms for topic 2.

Table 1: Counts of documents of each topic in each of the generated clusters using LDA model from gensim bow_corpus

V-measure score: 0.0049260284625080

topic	cluster	count
anger	0	312
anger	1	319
anger	2	386
anger	3	330
anger	4	415
anger	5	397
fear	0	308
fear	1	352

Table 2: Counts of documents of each topic in each of the generated clusters using LDA model from corpus_tfid

V-measure score: 0.0080090191811242

topic	cluster	count
anger	0	382
anger	1	446
anger	2	305
anger	3	336
anger	4	311
anger	5	379
fear	0	432
fear	1	342

fear	2	254
fear	3	400
fear	4	349
fear	5	274
joy	0	1030
joy	1	883
joy	2	975
joy	3	933
joy	4	733
joy	5	808
love	0	175
love	1	282
love	2	241
love	3	178
love	4	246
love	5	182
sadness	0	951
sadness	1	622
sadness	2	698
sadness	3	756
sadness	4	830
sadness	5	808
surprise	0	73
surprise	1	111
surprise	2	106
surprise	3	82
surprise	4	92
surprise	5	108

fear	2	280
fear	3	308
fear	4	323
fear	5	252
joy	0	805
joy	1	690
joy	2	914
joy	3	889
joy	4	1065
joy	5	999
love	0	245
love	1	108
love	2	163
love	3	226
love	4	302
love	5	260
sadness	0	781
sadness	1	842
sadness	2	721
sadness	3	922
sadness	4	632
sadness	5	767
surprise	0	44
surprise	1	108
surprise	2	146
surprise	3	80
surprise	4	83
surprise	5	111

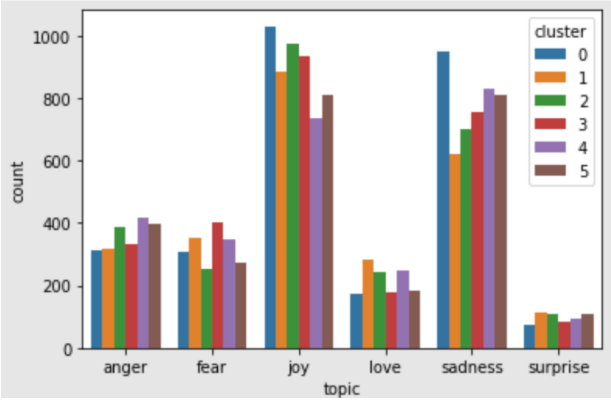


Figure 4: Plot of documents counts of each topic in each of the generated clusters using the LDA model from gensim bow_corpus

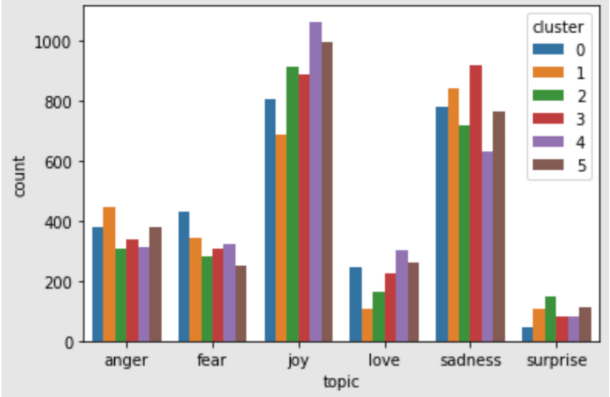


Figure 5: Plot of documents counts of each topic in each of the generated clusters using the LDA model from gensim bow_corpus

Appendix C HDP-LDA Output Visualizations

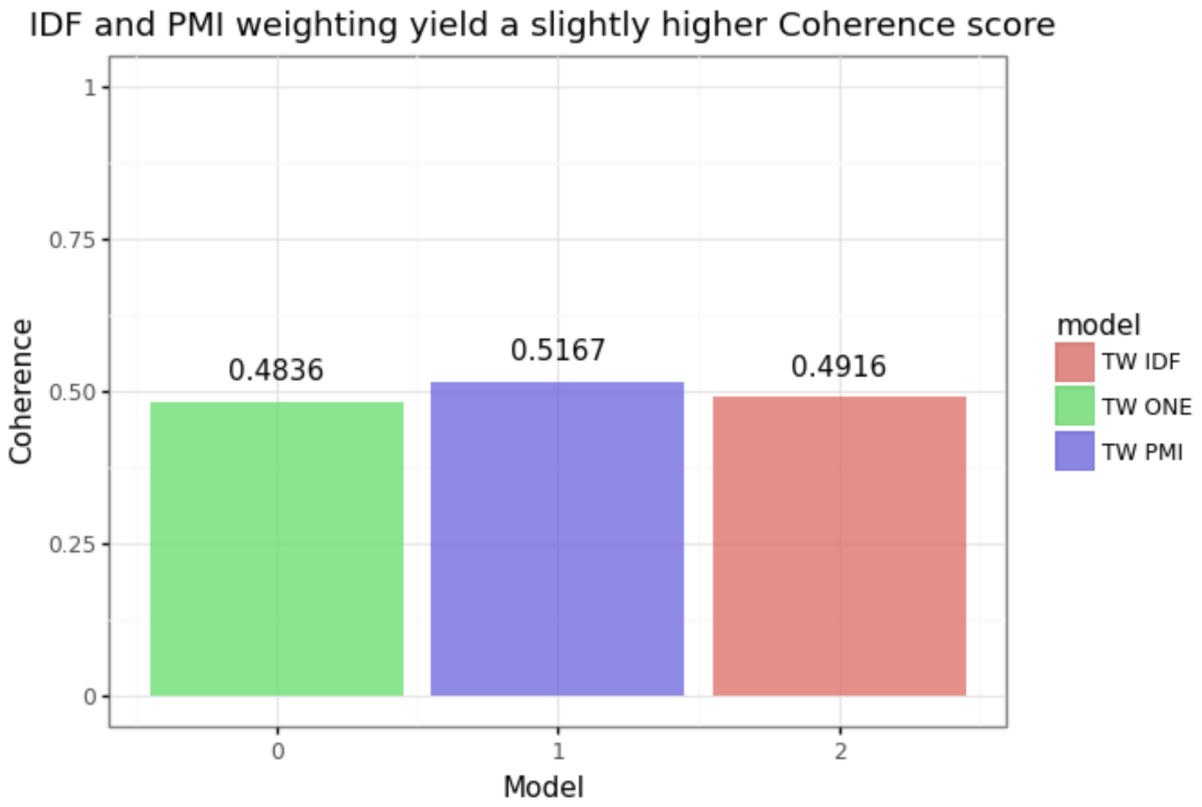


Figure 6: Coherence scores of how fit the word grouping are for a topic for the ONE, PMI, and IDF models

Appendix D Emotion Naive-Bayes Output Visualizations

Table 3: Counts of documents of each topic in each of the generated clusters using Naive-Bayes model

V-measure score: 0.5971668515814397

F1 score: 0.7806264577724985

topic	cluster	count
anger	0	15
anger	1	232
anger	2	5
anger	3	3
anger	4	13
anger	5	7
fear	0	17
fear	1	11
fear	2	2
fear	3	8
fear	4	182
fear	5	4
joy	0	14
joy	1	13
joy	2	58
joy	3	14
joy	4	11
joy	5	585
love	0	4
love	1	5
love	2	132
love	3	2

love	5	16
sadness	0	501
sadness	1	25
sadness	2	13
sadness	3	10
sadness	4	12
sadness	5	19
surprise	0	3
surprise	1	1
surprise	2	1
surprise	3	42
surprise	4	10
surprise	5	9

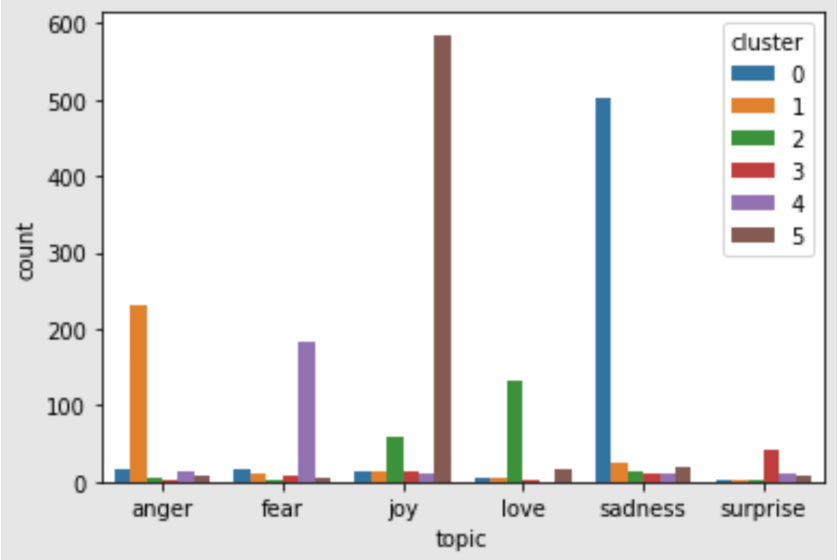


Figure 7: Plot of documents counts of each topic in each of the generated clusters using the Naive-Bayes model

Table 4-7: Counts of documents of each topic in each of the generated clusters using semi-supervised Naive-Bayes model. From left to right, the models are as follows: Soft EM on the training data set, Soft EM on the test data set, Hard EM on the training dataset, Hard EM on the test dataset. The V-measure scores from left to right are: 0.7669671920035445,

0.5700987286997194, 0.777863080032788, 0.5893649307648581. The F1 scores are: 0.8880947429311116, 0.7329266095376865, 0.8937319788053512, 0.7429693433056285.

topic	cluster	count
anger	0	107
anger	1	1988
anger	2	1
anger	3	1
anger	4	36
anger	5	26
fear	0	114
fear	1	58
fear	2	2
fear	3	5
fear	4	1709
fear	5	49
joy	0	111
joy	1	19
joy	2	52
joy	3	12
joy	4	47
joy	5	5121
love	0	19
love	1	6
love	2	1035
love	4	8
love	5	236
sadness	0	4491
sadness	1	34

topic	cluster	count
anger	0	26
anger	1	217
anger	2	1
anger	3	1
anger	4	17
anger	5	13
fear	0	28
fear	1	8
fear	2	2
fear	3	1
fear	4	175
fear	5	10
joy	0	16
joy	1	4
joy	2	22
joy	3	3
joy	4	12
joy	5	638
love	0	13
love	1	6
love	2	87
love	4	2
love	5	51
sadness	0	517
sadness	1	11

topic	cluster	count
anger	0	109
anger	1	1978
anger	2	3
anger	3	1
anger	4	37
anger	5	31
fear	0	80
fear	1	40
fear	2	2
fear	3	4
fear	4	1777
fear	5	34
joy	0	124
joy	1	27
joy	2	61
joy	3	6
joy	4	24
joy	5	5120
love	0	20
love	1	7
love	2	1042
love	4	3
love	5	232
sadness	0	4496
sadness	1	39

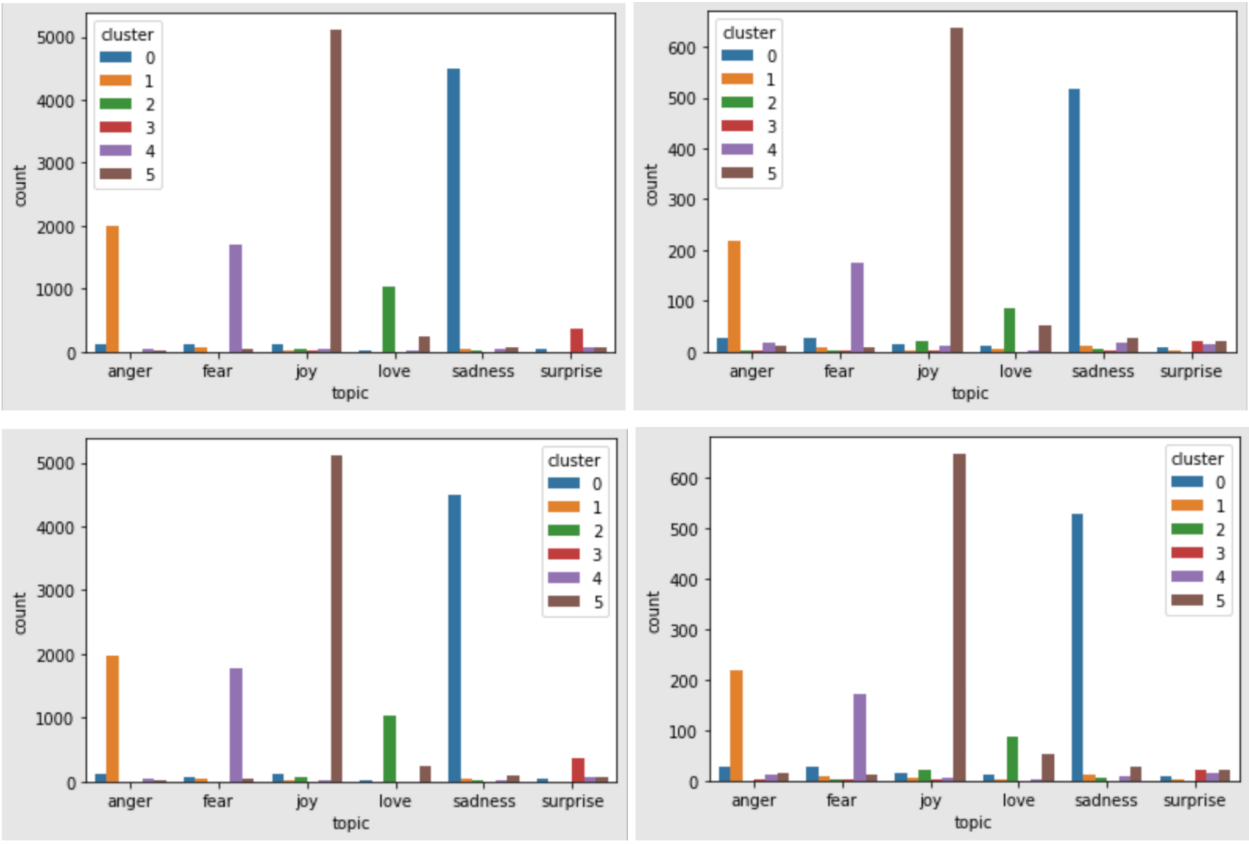
topic	cluster	count
anger	0	28
anger	1	218
anger	3	1
anger	4	13
anger	5	15
fear	0	28
fear	1	9
fear	2	1
fear	3	1
fear	4	172
fear	5	13
joy	0	14
joy	1	5
joy	2	22
joy	3	1
joy	4	6
joy	5	647
love	0	13
love	1	4
love	2	88
love	4	1
love	5	53
sadness	0	526
sadness	1	12
sadness	2	6

sadness	2	10
sadness	3	7
sadness	4	47
sadness	5	76
surprise	0	54
surprise	1	7
surprise	3	377
surprise	4	74
surprise	5	60

sadness	2	5
sadness	3	1
sadness	4	18
sadness	5	28
surprise	0	9
surprise	1	1
surprise	3	20
surprise	4	16
surprise	5	20

sadness	2	11
sadness	3	5
sadness	4	28
sadness	5	86
surprise	0	46
surprise	1	3
surprise	2	1
surprise	3	375
surprise	4	81
surprise	5	66

sadness	4	9
sadness	5	27
surprise	0	9
surprise	1	2
surprise	3	20
surprise	4	15
surprise	5	20



Figures 8-11 (top-left to bottom-right): Plot of documents counts of each topic in each of the generated clusters using the Naive-Bayes model with Soft EM on the training and test datasets (top row) and Hard EM on the training and test datasets (bottom row)