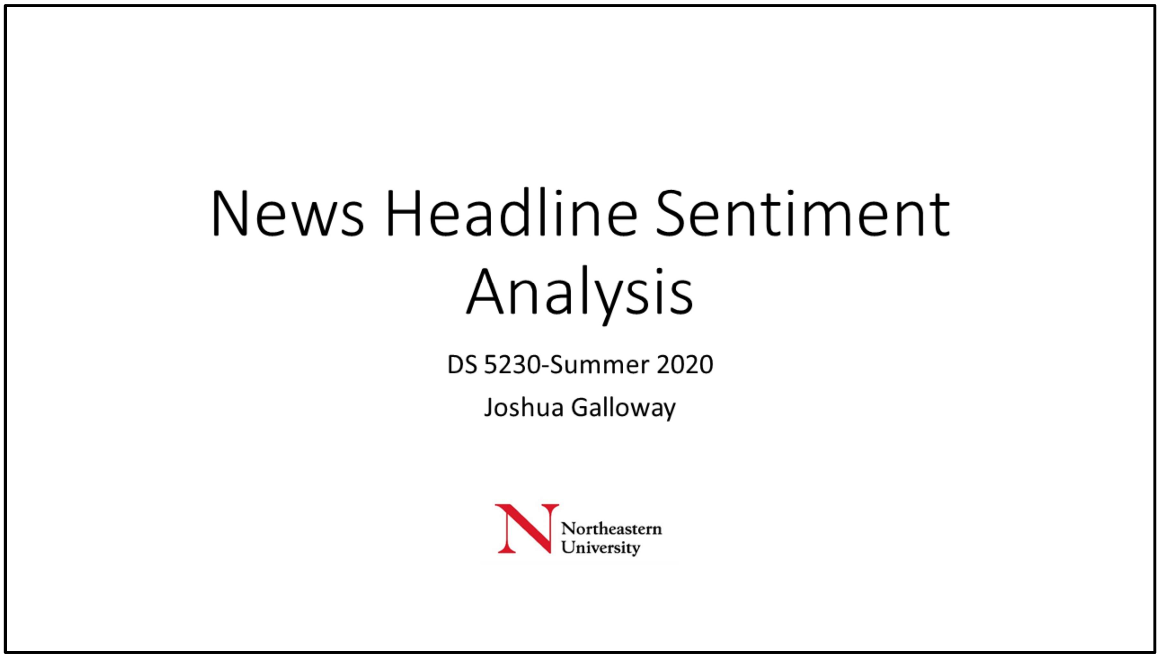


News Headline Sentiment Analysis

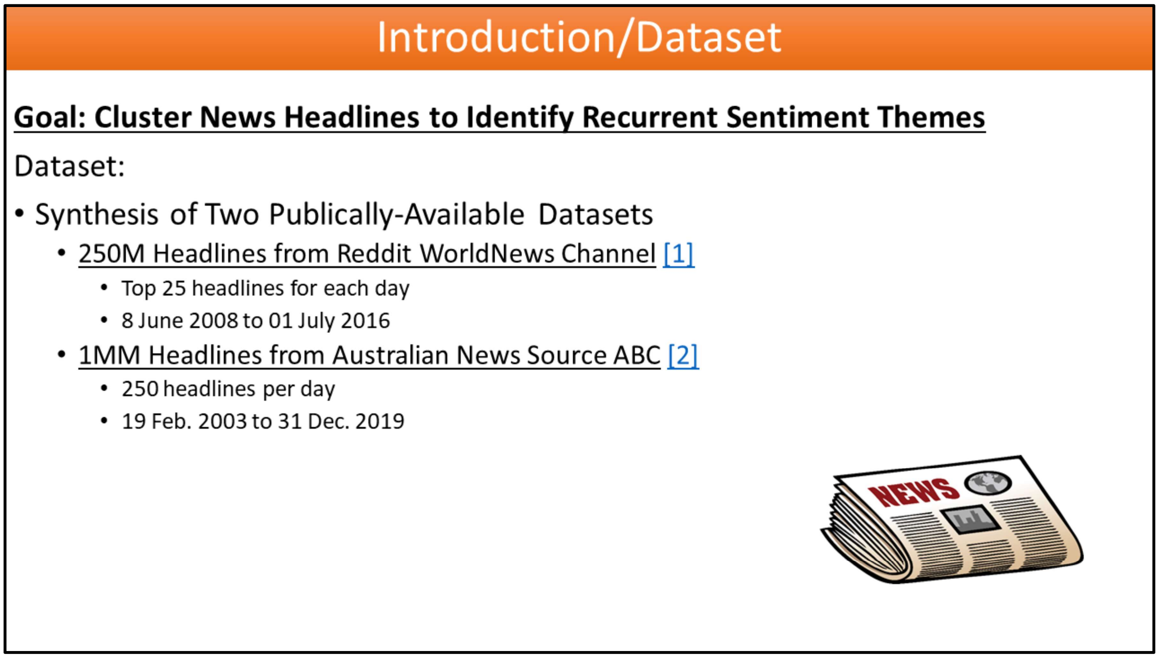
DS 5230-Summer 2020

Joshua Galloway

Northeastern University

News Headline Sentiment Analysis

- Introduction/Dataset
- Data Representation and Processing
- Model Derivation
- Results
- Conclusion



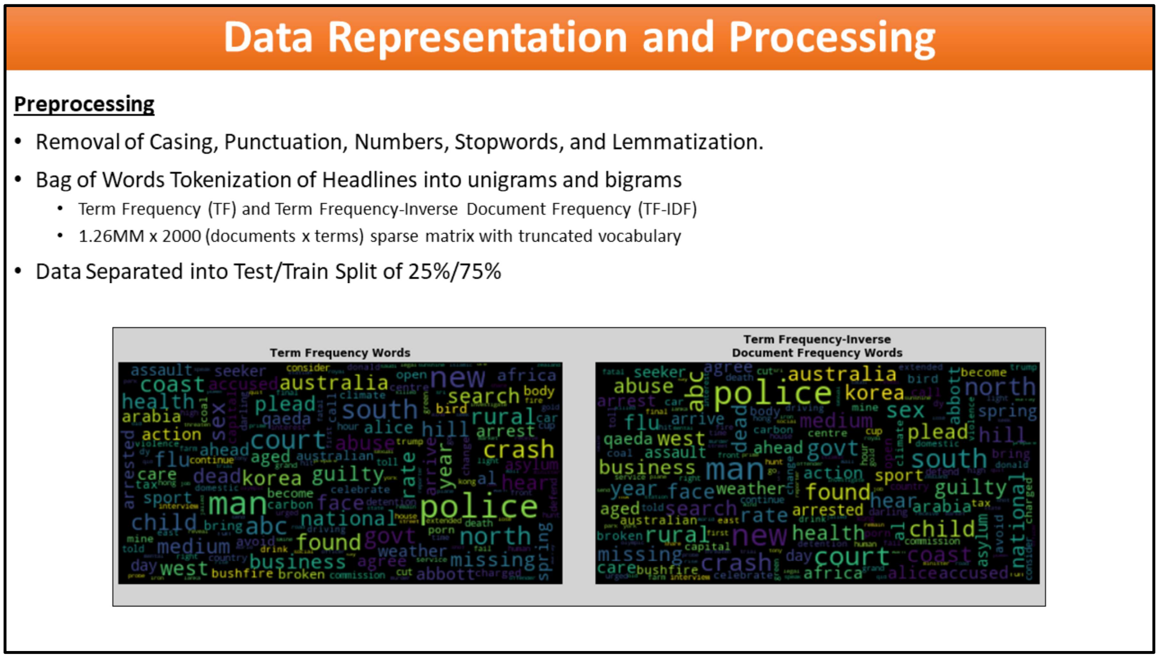
Introduction/Dataset

Goal: Cluster News Headlines to Identify Recurrent Sentiment Themes

Dataset:

- Synthesis of Two Publically-Available Datasets
  - 250M Headlines from Reddit WorldNews Channel [1]
    - Top 25 headlines for each day
    - 8 June 2008 to 01 July 2016
  - 1MM Headlines from Australian News Source ABC [2]
    - 250 headlines per day
    - 19 Feb. 2003 to 31 Dec. 2019

The goal of the project was to analyze a collection of news headlines to identify reoccurring sentiment themes. The dataset was comprised of roughly 1.25 million headlines and is the synthesis of two datasets publically available on Kaggle. The news sources from which the headlines were pulled were Reddit's WorldNews channel and ABC which is an Australian news organization. The time span for the collective dataset was roughly from 2003 to 2019.



Data Representation and Processing

Preprocessing

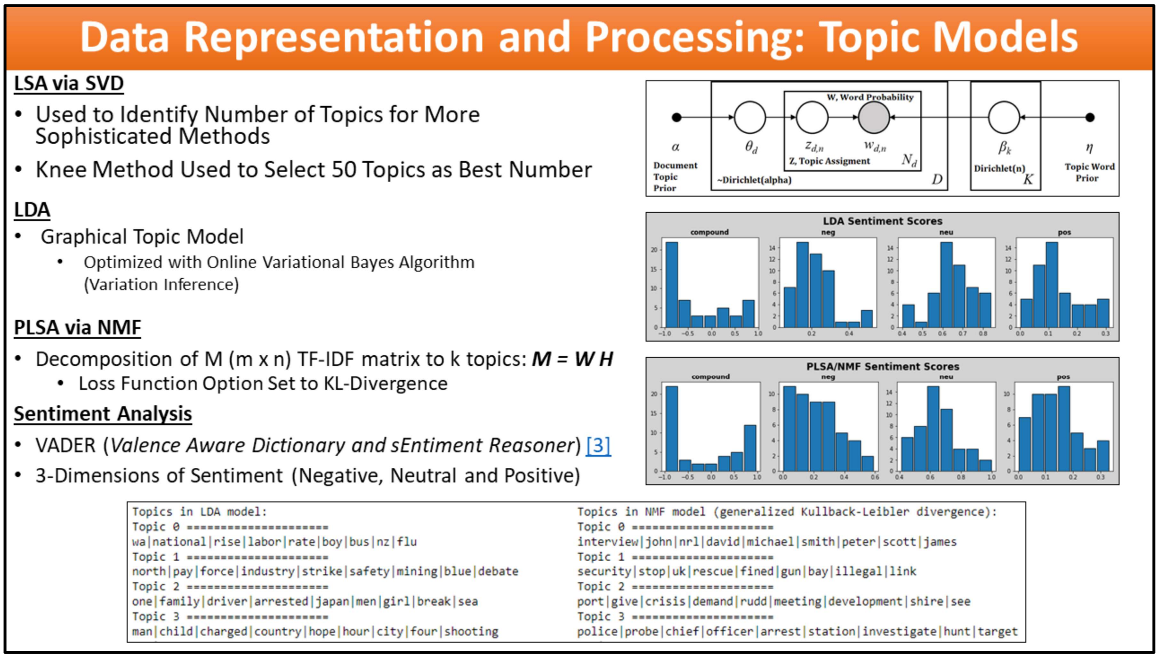
- Removal of Casing, Punctuation, Numbers, Stopwords, and Lemmatization.
- Bag of Words Tokenization of Headlines into unigrams and bigrams
  - Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF)
  - 1.26MM x 2000 (documents x terms) sparse matrix with truncated vocabulary
- Data Separated into Test/Train Split of 25%/75%

Term Frequency Words

Term Frequency-Inverse Document Frequency Words

Preprocessing the collection of headlines, which may also be called a corpus, was an intensive exercise. For each headline, the words had to be altered to aid in the tokenization and counting process. Casing, punctuation and numbers were all removed. Then stopwords, such as 'the', 'is', 'at' and so on, were removed. Following this, the remaining words were shortened to their root form. Two methods are readily used to accomplish this conversion--stemming and lemmatization. Stemming results in a word not necessarily found in natural language and so would not work with most available sentiment analyzers. For this reason, lemmatization was the chosen method whereby the words were reduced to their root. This is done to keep from producing additional terms which would not add value to the overall sentiment analysis such as a separate term for 'talk' and 'talking'.

The corpus was then tokenized in to unigrams and bigrams which loses the meaning associated with word order barring at most two-word phrases. This is called a bag of words model. Two models, or matrices, were produced—a Term Frequency (TF) matrix which is just a raw count of the words in each documents, and a Term Frequency-Inverse Document Frequency (TF-IDF) matrix which decreases the weighting of common words found in the corpus. These matrices were very large. Even with a vocabulary restrained to 2000 terms, the matrix was comprised of over 2 billion elements. These sparse matrices were separated after preprocessing into a test-train split of 25%/75% stratified by news source.



Data Representation and Processing: Topic Models

LSA via SVD

- Used to Identify Number of Topics for More Sophisticated Methods
- Knee Method Used to Select 50 Topics as Best Number

LDA

- Graphical Topic Model
  - Optimized with Online Variational Bayes Algorithm (Variation Inference)

PLSA via NMF

- Decomposition of M (m x n) TF-IDF matrix to k topics: ***M* = *W H***
  - Loss Function Option Set to KL-Divergence

Sentiment Analysis

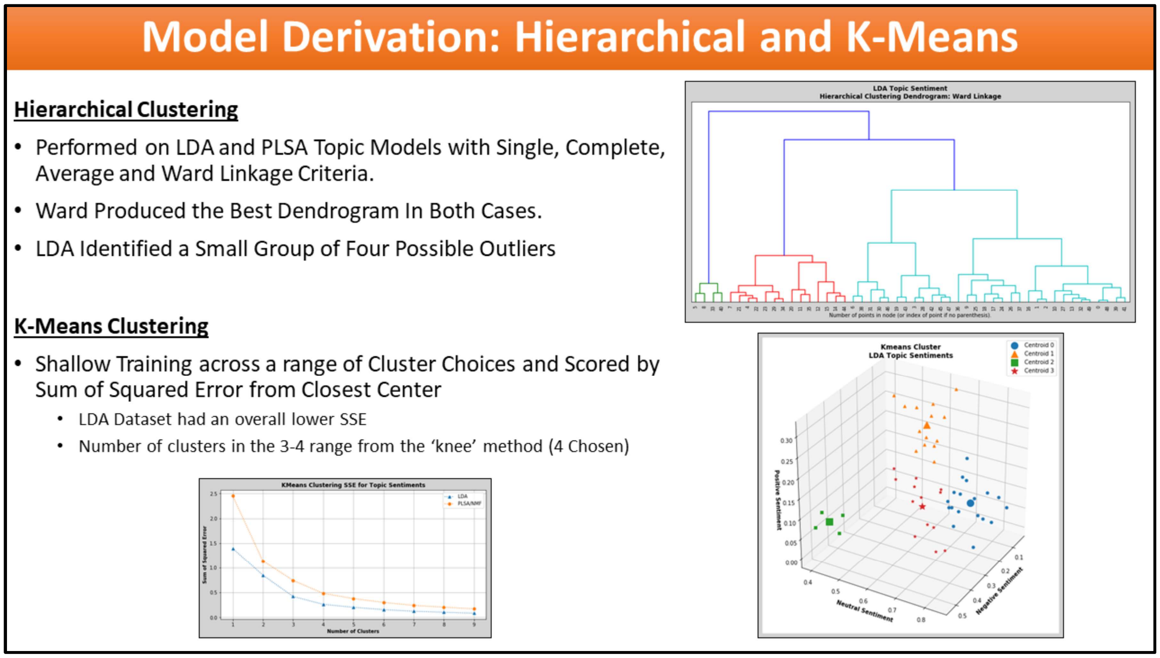
- VADER (*Valence Aware Dictionary and sEntiment Reasoner*) [3]
- 3-Dimensions of Sentiment (Negative, Neutral and Positive)

```
Topics in LDA model:                                          Topics in NMF model (generalized Kullback-Leibler divergence):
Topic 0 =====================                                 Topic 0 =====================
wa|national|rise|labor|rate|boy|bus|nz|flu                    interview|john|nrl|david|michael|smith|peter|scott|james
Topic 1 =====================                                 Topic 1 =====================
north|pay|force|industry|strike|safety|mining|blue|debate     security|stop|uk|rescue|fined|gun|bay|illegal|link
Topic 2 =====================                                 Topic 2 =====================
one|family|driver|arrested|japan|men|girl|break|sea           port|give|crisis|demand|rudd|meeting|development|shire|see
Topic 3 =====================                                 Topic 3 =====================
man|child|charged|country|hope|hour|city|four|shooting        police|probe|chief|officer|arrest|station|investigate|hunt|target
```

Since the size of the processed dataset was so large, dimensional reduction was necessary in order to begin the analysis. Principle component analysis (PCA) and t-distributed stochastic neighbor embedding (T-SNE) were tried but failed due to the size of the matrices involved, as they would not fit in memory. So, Topic Modeling was used as a form of dimensional reduction. Latent Semantic Analysis (LSA) was performed via singular value decomposition (SVD) on the term frequency matrix. This is less memory intensive than PCA because the covariance matrix need not be calculated. Examining the Singular Values, a knee appears around 50 topics, and that number was selected for the number of Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) topics to model.

LDA is a graphical model in which word document count is the observed variable and the topic word prior, document topic prior and document to topic assignment prior are latent (or hidden) variables. In sklearn's implementation the model is fit by maximizing the Evidence Lower Bound which is equivalent to minimizing the KL-divergence. LDA was performed on the term frequency matrix. PLSA is similar to LSA but the decomposition is performed by non-negative matrix factorization (NMF) with loss function set to KL divergence on the TF-IDF matrix [4].

Sentiment Analysis was carried out with the VADER sentiment analysis lexicon. This lexicon produces four metrics in the form of a [-1,1] overall score (negative to positive sentiment), and three [0,1] metrics for gradations of negative, neutral and positive sentiment. Examining the sentiment scored topics from each model shows that the sentiment distributions are similar, with the PLSA model being slightly more positive in general than the LDA topic model.



Model Derivation: Hierarchical and K-Means

Hierarchical Clustering

- Performed on LDA and PLSA Topic Models with Single, Complete, Average and Ward Linkage Criteria.
- Ward Produced the Best Dendrogram In Both Cases.
- LDA Identified a Small Group of Four Possible Outliers

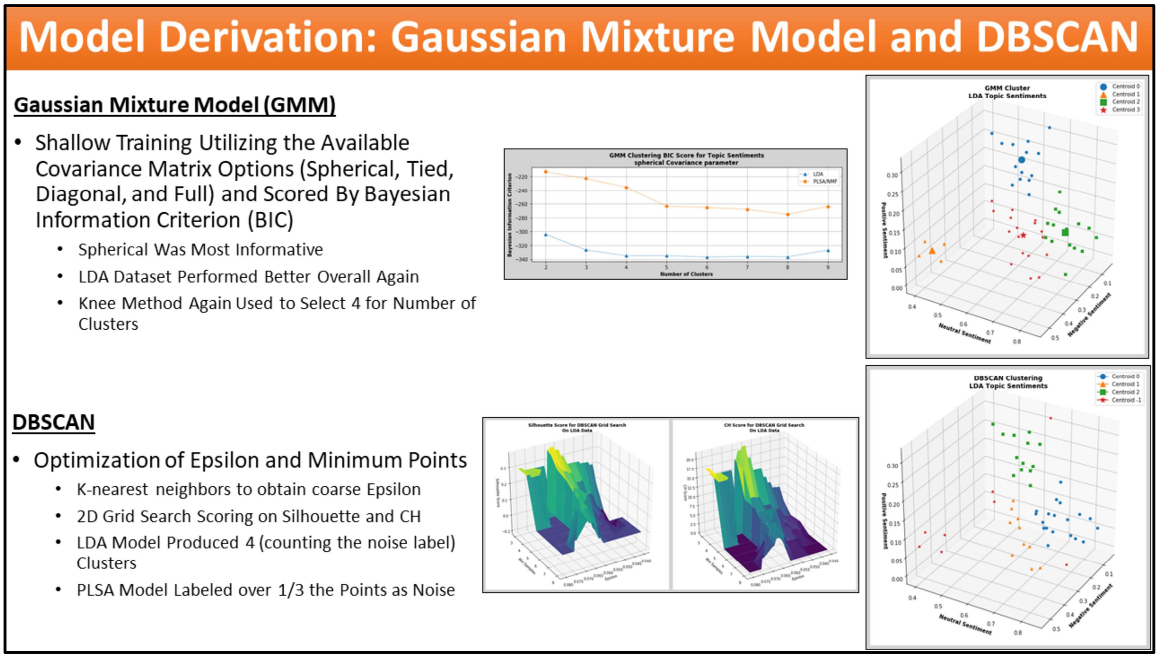
K-Means Clustering

- Shallow Training across a range of Cluster Choices and Scored by Sum of Squared Error from Closest Center
  - LDA Dataset had an overall lower SSE
  - Number of clusters in the 3-4 range from the ‘knee’ method (4 Chosen)

Hierarchical Clustering was the first means of clustering analysis on the topic models. The four available linkage metrics in sklearn’s implementation were used to cluster both reduced datasets based on the 3-dimensional sentiment scoring. Of the clustering criteria, Ward or intra-cluster variance linkage produced the best dendrogram in both cases. For the LDA dataset reoccurring themes appeared throughout all four metrics. A small group of 4 outliers, seen in green in the dendrogram, were identifiable, and an overall cluster number of four seemed appropriate with all linkage metrics.

With respect to the PLSA dataset, again four clusters seemed the best choice from the useful dendrograms. The single link criterion did not produce serviceable results as the data was clustered into long thin groupings likely due to the closest pair grouping behavior designed into the metric.

K-Means was tried next, and the knee method used to find the optimal number of clusters. The graph of sum of squared error versus number of clusters shows that the LDA dataset performs best overall and that the optimal number of clusters is in the 3-4 range. Imbuing the findings of the hierarchical clustering exercise, the model was trained more deeply with the same cluster numbers selection of four.



Model Derivation: Gaussian Mixture Model and DBSCAN

Gaussian Mixture Model (GMM)

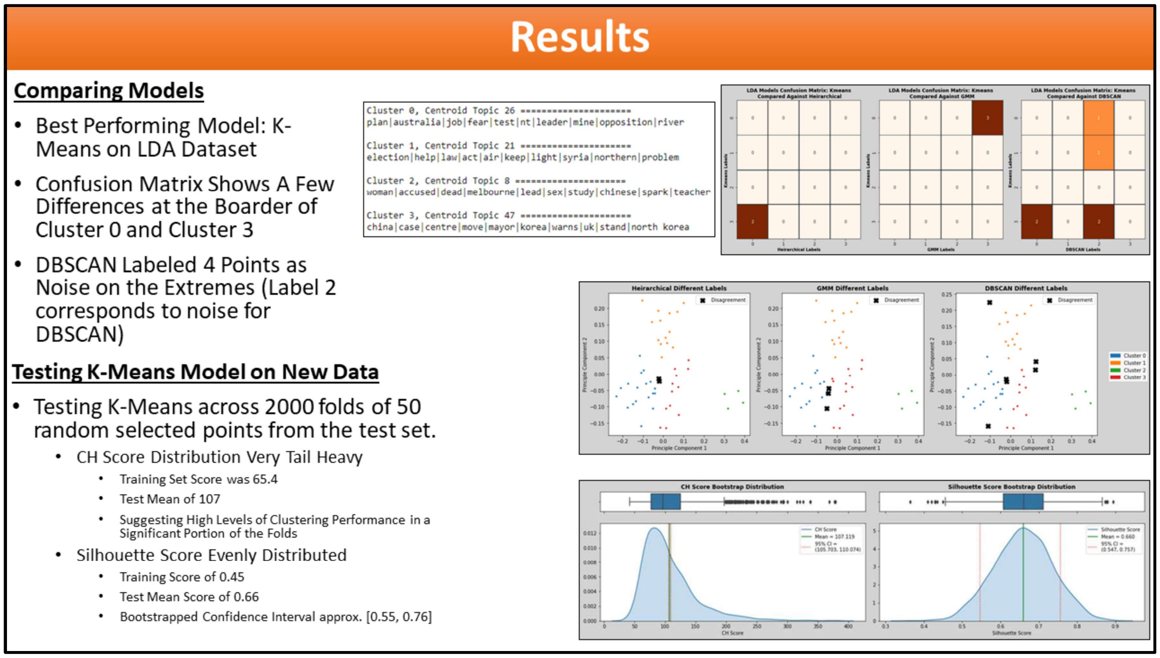
- Shallow Training Utilizing the Available Covariance Matrix Options (Spherical, Tied, Diagonal, and Full) and Scored By Bayesian Information Criterion (BIC)
  - Spherical Was Most Informative
  - LDA Dataset Performed Better Overall Again
  - Knee Method Again Used to Select 4 for Number of Clusters

DBSCAN

- Optimization of Epsilon and Minimum Points
  - K-nearest neighbors to obtain coarse Epsilon
  - 2D Grid Search Scoring on Silhouette and CH
  - LDA Model Produced 4 (counting the noise label) Clusters
  - PLSA Model Labeled over 1/3 the Points as Noise

Next a set of Gaussian Mixture Models (GMM) were built. For all of the available covariance matrix options, the model was shallowly trained with varied numbers of clusters and scored on Bayesian Information Criterion. Of the effective models, spherical performed the best and again the LDA model scored better that the PLSA model. The knee on the LDA curve suggests that again four is the best number of clusters, and the model was more deeply trained with this knowledge.

Finally, Density-Based Spatial Clustering and Application with Noise (DBSCAN) was tried. Again utilizing the knee method, k-nearest neighbors was used to find a rough value of epsilon which is the radius around each point defining the neighborhood for this algorithm. From this starting point, a two-dimensional grid search in minimum points per cluster and epsilon scoring on both the Silhouette coefficient and CH score was run. With the optimized parameters, models were trained and again scored on Silhouette coefficient and CH score with the noisy points identified by the algorithm removed. In this case, the PLSA-based model out performed the LDA-based model and also differed in the number of clusters the algorithm identified. The LDA model produced 3 overall clusters (including the noisy points as its own cluster) with 16% of the points labeled as noise. The PLSA dataset found 4 clusters and labeled 36% of the points as noise making the results of the scoring slightly suspect.



Results

Comparing Models

- Best Performing Model: K-Means on LDA Dataset
- Confusion Matrix Shows A Few Differences at the Boarder of Cluster 0 and Cluster 3
- DBSCAN Labeled 4 Points as Noise on the Extremes (Label 2 corresponds to noise for DBSCAN)

Testing K-Means Model on New Data

- Testing K-Means across 2000 folds of 50 random selected points from the test set.
    - CH Score Distribution Very Tail Heavy
        - Training Set Score was 65.4
        - Test Mean of 107
        - Suggesting High Levels of Clustering Performance in a Significant Portion of the Folds
    - Silhouette Score Evenly Distributed
        - Training Score of 0.45
        - Test Mean Score of 0.66
        - Bootstrapped Confidence Interval approx. [0.55, 0.76]

```
Cluster 0, Centroid Topic 26 ======================
plan|australia|job|fear|test|nt|leader|mine|opposition|river

Cluster 1, Centroid Topic 21 ======================
election|help|law|act|air|keep|light|syria|northern|problem

Cluster 2, Centroid Topic 8 ======================
woman|accused|dead|melbourne|lead|sex|study|chinese|spark|teacher

Cluster 3, Centroid Topic 47 ======================
china|case|centre|move|mayor|korea|warns|uk|stand|north korea
```

LDA Models Confusion Matrix: Kmeans Compared Against Heirarchical

Kmeans Labels \ Heirarchical Labels	0	1	2	3
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	2	0	0	0

LDA Models Confusion Matrix: Kmeans Compared Against GMM

Kmeans Labels \ GMM Labels	0	1	2	3
0	0	0	0	3
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0

LDA Models Confusion Matrix: Kmeans Compared Against DBSCAN

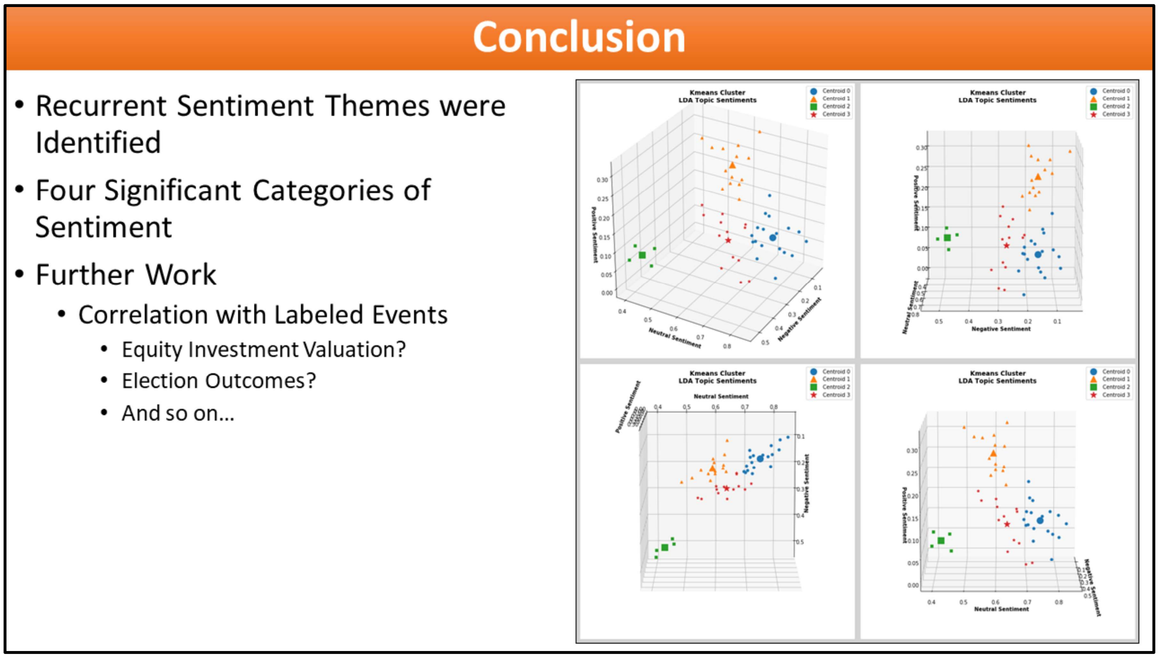
Kmeans Labels \ DBSCAN Labels	0	1	2	3
0	0	0	1	0
1	0	0	1	0
2	0	0	0	0
3	2	0	2	0

The K-means LDA-based model performed best based on Silhouette coefficient and CH score metrics. In fact, the LDA-based models performed better in all cases with the exception of DBSCAN. This is likely due to the additional sophistication in the LDA algorithm from the inclusion of extra latent variables versus the non-negative matrix factorization.

Comparing the models in a confusion matrix shows only a few labels were in disagreement. Inspecting these disagreements on a PCA-reduced two dimensional graphing of the LDA dataset clarifies the region of disagreement for all the algorithms is generally between the boundary of cluster 0 and 3. Inspecting the topics closest to the Cluster 0 and 3 centroids shows that the two topics are slightly similar with the words ‘fear’ and ‘warns’ and other vaguely worrisome words associated with each topic.

The reserved test set was utilized to measure the performance of the model. Using only words from the model’s 2000 term vocabulary, the headlines in the test set were scored for sentiment; following which, the K-Means model was used to predict a label for each headline. Each randomly selected 50-point testing fold was scored on Silhouette coefficient and CH score. The CH score distribution was very tail heavy showing high levels of clustering performance for a significant portion of the folds. The box plot for the CH score distribution shows many outliers to the higher performing side of the interquartile range. The Silhouette coefficient distribution was shown to be roughly normal with a mean score of 0.66 and 95% confidence interval from 0.55 to 0.76.

In both cases, the scoring was improved on the test set over the training score. This is likely due to duplication of topics between headlines creating tighter clusters with more distance between them, as this would improve either metric.

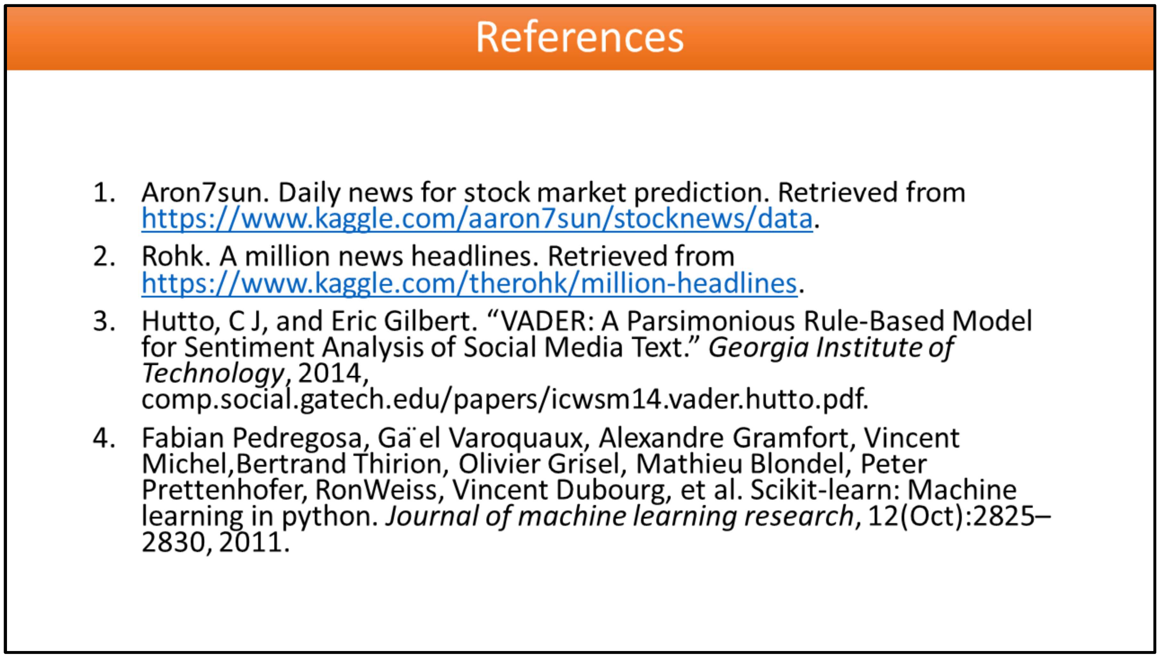


Conclusion

- Recurrent Sentiment Themes were Identified
- Four Significant Categories of Sentiment
- Further Work
  - Correlation with Labeled Events
    - Equity Investment Valuation?
    - Election Outcomes?
    - And so on…

Overall, the data seems to support the idea that news headlines may be segmented into a set of recurrent sentiment themes. Since only DBSCAN's LDA model was in disagreement in optimal cluster numbers between algorithms, the data also supports that headlines should be segregated in to 4 major sentiment categories.

Further work from here may be to collect a more diverse dataset and try a varied number of sentiment analyzers. Work could also include adding a temporal dimension and correlating the clustering with measured events, such as equity investment pricing or election outcomes, in order to use the information to predict future outcomes.



References

1. Aron7sun. Daily news for stock market prediction. Retrieved from https://www.kaggle.com/aaron7sun/stocknews/data.
2. Rohk. A million news headlines. Retrieved from https://www.kaggle.com/therohk/million-headlines.
3. Hutto, C J, and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Georgia Institute of Technology*, 2014, comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf.
4. Fabian Pedregosa, Gä el Varoquaux, Alexandre Gramfort, Vincent Michel,Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, RonWeiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.