Methods description

* Briefly describe your unsupervised learning workflow, the learning methods you used, and the feature representations you chose. You must justify why you chose your methods.
* You should have an adequate number and nature of methods: a minimum of two unsupervised methods must be described and explored (i.e. with very different underlying mechanisms, e.g. probabilistic, non-probabilistic, tree-based, instance-based, etc).  This might be variable with larger or smaller teams and project specifics, with instructor permission.
* Include a description of how you did hyperparameter tuning or exploration with your models.
* Methods used must be clearly described, each with correct justification.
  1. **LDA Topic Modeling on Consumer Complaints Narrative:**

Workflow:

* Selected only those rows where the 'Consumer complaint narrative' was available.
* Preprocessed the narratives by:
* Converting to lowercase.
* Removing masking patterns like XX, XXX, XXXX, …
* Removing punctuation, numbers, and stopwords.
* Removing extra spaces.
* Converted the cleaned narratives into a bag-of-words representation.
* Used the LDA (Latent Dirichlet Allocation) method to identify topics within these narratives.
* Visualized the top topics using word clouds.

Justification:

LDA is a popular method for topic modeling which helps in understanding the hidden thematic structure in a large collection of texts. Here, it helps in understanding the major concerns or issues that consumers have.

* 1. **Association rule mining, association rules between the product and issue:**

Workflow:

* Focused on the 'Product' and 'Issue' columns.
* One-hot encoded the columns to prepare the data for association rule mining.
* Used the Apriori algorithm to find frequent itemsets.
* Generated association rules based on the frequent itemsets.
* Visualized the top association rules using network graphs.

Justification:

Association rule mining, especially the Apriori algorithm, is suitable for finding interesting relationships or associations between variables. Here, it helps in understanding which products are associated with which issue, uncovers relationships between products and issues that may not be immediately apparent. It allows organizations to spur investigations into why certain product-issue combinations are frequent.

* 1. **Finding Similar Complaints with Word2Vec and Cosine Similarity:**

Workflow:

* Preprocessed the complaint narratives.
* Loaded a pre-trained Google Word2Vec model.
* Computed the average Word2Vec vector for each complaint narrative.
* Project each narrative onto 300D vector space.
* Computed the cosine similarity between these vectors to find similar complaints.

Justification:

Word2Vec captures semantic information about words. By computing the average Word2Vec vector for each complaint, we capture the semantic essence of each complaint narrative. Cosine similarity then helps in finding complaints that are semantically similar. The implementation matches new entry to similar past complaints for quick resolution. It reduces manual effort in searching past records for similar cases, thus improves customer satisfaction by providing consistent responses.

* 1. **Clustering Companies based on Complaints with TF-IDF and K-mean:**

Workflow:

* Combined all narratives for each company into a single text.
* Converted these combined narratives into a TF-IDF matrix.
* Used the elbow method to determine the optimal number of clusters.
* Used K-means clustering to cluster the companies based on their TF-IDF representations.

Justification:

K-means clustering helps in grouping data points (in this case, companies) that have similar features. By using TF-IDF representations, we are clustering companies that receive similar types of complaints. This helps in identifying companies with similar operational issues. It enables targeted actions for each company cluster instead of one-size-fits-all approach. It focuses improvement efforts on clusters with most severe or frequent complaint types.

**Hyperparameter tuning / Exploration**

For LDA topic modeling, the number of topics is a key hyperparameter that can greatly impact the resulting topics. The coherence score metric was used to tune the number of topics. Models were trained with different numbers of topics (5, 10, 15, 20). The one with the highest coherence score was selected, which was 5 topics.

For K-Means clustering, the elbow method was used to tune the number of clusters k. K-Means models were trained with different k values ranging from 1 to 30 clusters. The inertia was plotted for each k, and the "elbow point" was selected where inertia did not decrease much with more clusters. This led to selecting 14 clusters.

**Method categories**

Probabilistic Methods: LDA topic modeling, which assumes that documents are mixtures of topics and that topics have a probability distribution over words.

Instance-based Methods: Association rule mining and K-means clustering, where we are directly working with individual instances (complaints or companies) to discover patterns or clusters.

* Unsupervised Evaluation
  + In this section you will provide a correct and comprehensive evaluation, analyzing the effectiveness of both your methods, and your choice of feature representation.
  + Overall results reporting.
    - State and justify your choice of evaluation metrics used.
    - Provide at least one overall summary of results that compares the best model from each family you used, in a clear, concise table.
    - To summarize your findings, include at least two visualizations (chart, plot, tag cloud, map or other graphic) for each unsupervised method used that summarize your analysis.
  + (5 points) Do at least one sensitivity analysis on your best model: How sensitive are your results to choice of (hyper-)parameters, features, or other varying solution elements?

State and justify your choice of evaluation metrics used.

Evaluation metrics used for each unsupervised learning method and justification:

1. **LDA Topic Modeling:**

* Coherence score: Evaluates semantic interpretability of topics. Helps tune number of topics.
* Topic distributions: Useful for exploring relative prevalence of topics.

Justification: Coherence is a standard metric for topic model evaluation. Topic distributions give insights into data.

1. **Apriori Association Rules**

Support: Frequency of occurrence of a rule. Filters rare rules.

Confidence: Reliability of a rule. Filters spurious rules.

Lift: Interestingness of a rule, ratios between expected and observed support. Prioritizes useful rules.

Justification: Support, confidence and lift are standard metrics for filtering and ranking association rules.

1. **Word2Vec Similarity**

* Cosine similarity: Measures how similar two mean vectors are based on their orientation. Gives semantic similarity of text.

Justification: Cosine similarity is the standard metric used with Word2Vec models for comparing word/document vectors.

1. **K-Means Clustering**

* Inertia: Sum of squared distances between data points and cluster centroids. Lower is better.
* Cluster counts: Number of data points assigned to each cluster. Indicates relative sizes.

Justification: Inertia helps determine optimal number of clusters. Inertia measures compactness of clustering. The Elbow Method finds the "sweet spot" where adding more clusters does not significantly improve inertia. It balances between too few and too many clusters.

Cluster counts give insights into sizes and to check if there is any outlier.

Provide at least one overall summary of results that compares the best model from each family you used, in a clear, concise table.

|  |  |  |
| --- | --- | --- |
| **Method** | **Model Details** | **Key Insight** |
| Topic Modeling (LDA) | 5 topics, tuned by coherence score | Main consumer complaint topics identified with human-in-the-loop check (some topics mined are: credit report, debt collection, loan etc) |
| Association Rule Mining (Apriori) | Lift >= 8 | Strong associations between products and issues uncovered. We focus on strong positive association between two entities. The top ten are ranging from 8 – 28 |
| Document Embedding (Word2Vec) | Google News pre-trained embeddings | Semantic similarity between complaints calculated. Identified narrative meaning with human-in-the-loop check. |
| Clustering (K-Means) | 10 clusters, tuned by elbow method | Companies clustered by complaint patterns. E.g Cluster 1 are mostly banks related, while Cluster 5 covers a large portion of automotive companies. |

To summarize your findings, include at least two visualizations (chart, plot, tag cloud, map or other graphic) for each unsupervised method used that summarize your analysis.

1. LDA

A graph with blue lines

Description automatically generated

The visualization illustrates the relationship between the number of topics chosen for LDA topic modeling and the coherence score for those topics.

There's a noticeable initial increase in coherence as the number of topics goes up, peaking at 5 topics. Follow by a Subsequent Decline. This suggests that adding more topics beyond 6 results in less coherent or overlapping topics,

A graph showing a number of people

Description automatically generated with medium confidence

The visualization presents the distribution of topics and their corresponding weightage in the dataset of complaints. The terms within each topic offer insights into the nature of cluster the complaints.

1. Association rule

A screenshot of a document

Description automatically generated

The visualization presents a heatmap that depicts the association rules between various financial products and the issues related to them. The heatmap's colors correspond to the lift values, with darker shades indicating higher lift values. Lift values represent the likelihood of the association between the product and the issue.

A group of blue dots and lines

Description automatically generated

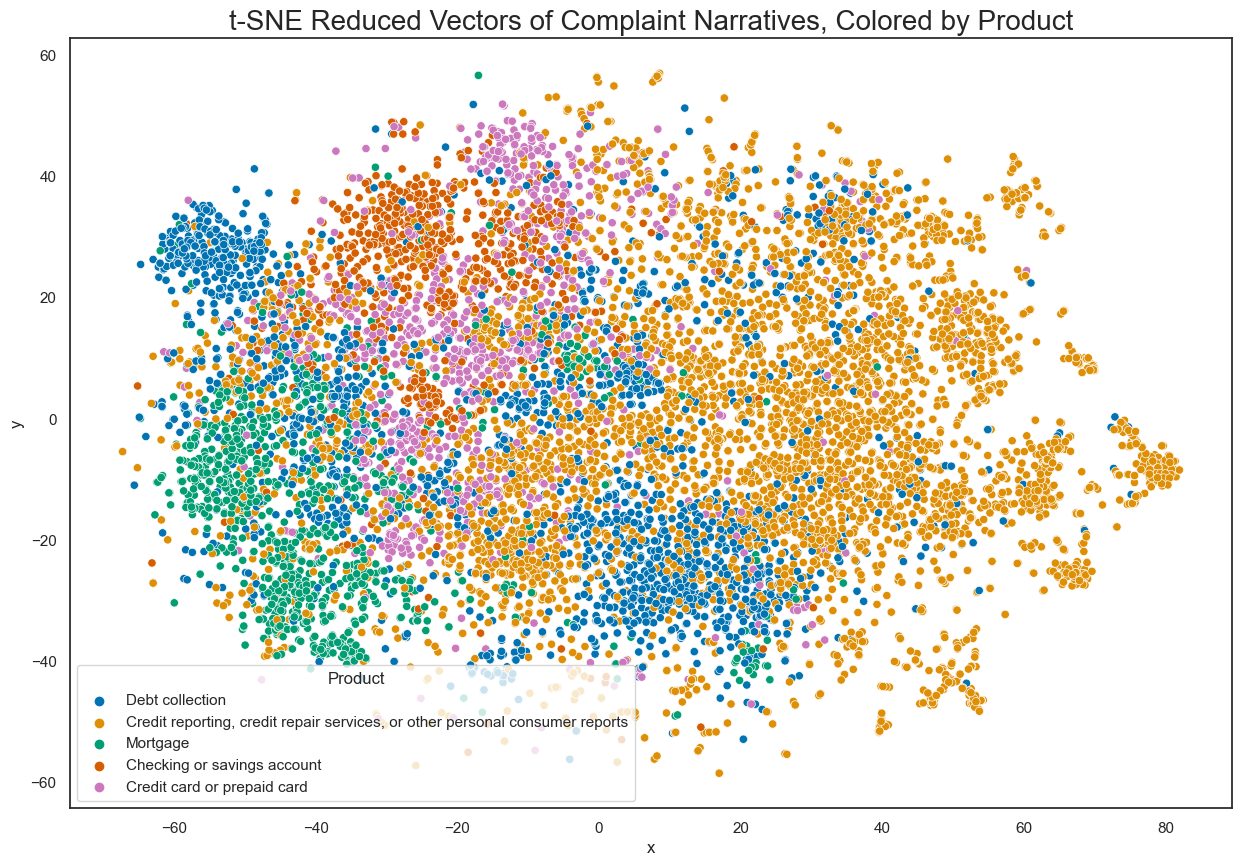
The visualization provides a network of associations between various products and the issues associated with them. The boldness of edge line represent the degree of lift metric. Certain products, such as those related to credit reporting and debt collection, have multiple issues associated with them, highlighting the complexity and range of challenges consumers might face in these areas.

1. Word2Vec Similarity

A graph of a person with a blue line

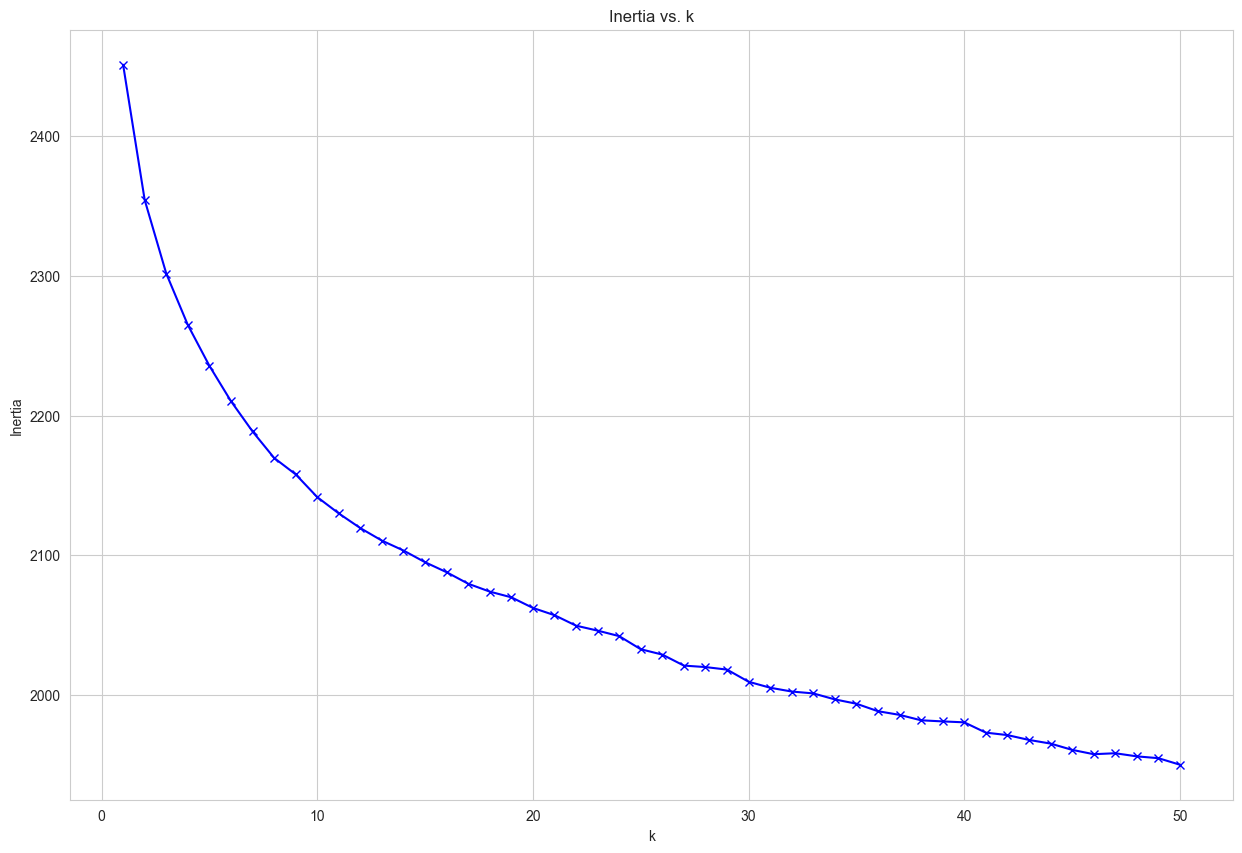
Description automatically generated with medium confidence

The visualization is a histogram that depicts the distribution of Cosine Similarity Scores with every possible pair of complaint narrative. In essence, the histogram shows that the majority of the data has high cosine similarity scores, centering around a mean value of 0.70.

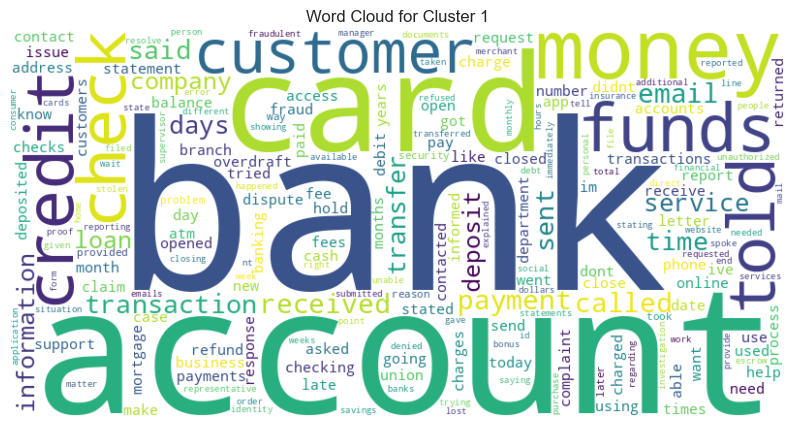


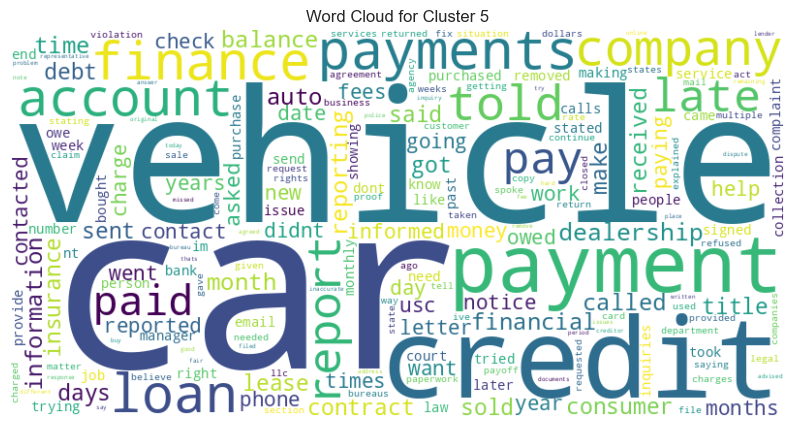
The visualization presents a scatter plot of t-SNE reduced vectors of complaint narratives. These points are colored based on top 5 product associated with each complaint.

1. Clustering



The visualization presents an "elbow plot" showing the relationship between the number of clusters (k) and the corresponding inertia. The key observation from such plots is to find the "elbow point", which is where the rate of decrease in inertia sharply changes, we think the sweet spot is between 10-20 in this case.





The visualization provides word clouds for two distinct clusters, Cluster 1 and Cluster 5. It helps on alignment of meaningfulness of the clusters with human perception.

Do at least one sensitivity analysis on your best model: How sensitive are your results to choice of (hyper-)parameters, features, or other varying solution elements?

A graph showing the value of a stock market

Description automatically generated

The relationship between the "max\_features" parameter of a TF-IDF vectorizer and the coherence score related to LDA topic modeling.

General Trend: The coherence score seems to vary as the "max\_features" parameter changes. This indicates that the number of features considered by the vectorizer affects the quality or coherence of the topics generated by LDA.

Initial Stability: For values of "max\_features" from about 1000 to around 4000, the coherence score remains relatively stable with minor fluctuations. This suggests that within this range, the topics' quality doesn't drastically change.

Peak Coherence: Around "max\_features" value of 5000, there is a sharp peak in the coherence score. This indicates that the LDA model achieved the best topic coherence with a vectorizer set to consider 5000 features. It suggests that this might be an optimal value for modeling, as the topics derived from the data are most coherent (or meaningful) with this setting.

Decrease after Peak: there's a rapid decrease in coherence as "max\_features" increases. This suggests that after a certain point, adding more features (terms) to the model starts to degrade the quality of the topics. It might indicate noise being introduced, which can dilute the significance of more relevant terms.

Sensitivity: The model appears to be most sensitive to the "max\_features" parameter around the 5000 mark. Before and after this point, the results either remain stable or degrade rapidly. It underscores the importance of parameter tuning in achieving optimal results.