Sentiment Analysis for Customer Feedback

Report on Data Cleaning and Exploratory Data Analysis

Date: June 2, 2025
Prepared by: Madhu Kumari
Mentor: Arockia Liborious

This week, I focused on data cleaning and exploratory data analysis (EDA) phase for sentiment analysis. The primary objective was to understand the structure, distribution, and characteristics of the text data, as well as to evaluate different sentiment analysis approaches.

Data Cleaning and Preprocessing

- **Text Normalization:** Convert all text to lowercase for consistency.
- **Remove Punctuation and Special Characters:** Clean out symbols and HTML tags.
- **Filter Stop Words:** Remove common words that do not contribute to sentiment (e.g., “the”, “is”).
- **Tokenization:** Break down sentences into individual words or tokens for easier processing.
- **Lemmatization:** Reduce words to their base forms (e.g., "running" becomes "run") to consolidate similar terms.
- **Optional Stemming:** Use stemming to quickly reduce words to their roots, though this may be less precise than lemmatization.
- **Number Handling:** Decide whether to keep or remove numbers, as they can sometimes impact sentiment in financial contexts.
- **Whitespace Management:** Remove extra spaces and newlines that may have been introduced during data collection.
- **Handle Contractions and Slang:** Expand contractions and normalize informal language.
- **Deduplication:** Remove duplicate feedback entries.

- **Language Filtering:** Ensure feedback is in the target language or use multilingual tools if needed.

Exploratory Data Analysis Techniques

Basic Text Statistics

- **Text Length Analysis:** Analyze character and word counts per feedback item using histograms or box plots.
- **Unique Word Count:** Determine vocabulary richness by counting unique words.
- **Frequency of Words:** Identify and visualize the most frequent words or phrases using bar charts or word clouds.

Sentiment Distribution Analysis

- **Apply Sentiment Analysis:** Use tools like VADER, TextBlob, or custom models to assign sentiment scores (positive, negative, neutral).
- **Visualize Sentiment:** Plot histograms or pie charts to show the distribution of sentiment across feedback.
- **Time Series Analysis:** If dates are available, track sentiment trends over time to spot patterns or changes.

Aspect-Based and Emotion Analysis

- **Aspect Identification:** Use NLP techniques to extract specific features or topics mentioned in feedback (e.g., product, service, delivery).
- **Emotion Detection:** Identify underlying emotions (anger, joy, frustration) expressed in feedback for deeper insights.
- **Visualize by Aspect/Emotion:** Create visualizations to show sentiment or emotion distribution for each aspect.

Class Imbalance and Outlier Detection

- **Label Distribution:** If feedback is labeled (e.g., by rating or sentiment), check for class imbalance using bar or pie charts.

- **Outlier Detection:** Identify unusually long or short feedback items, or extreme sentiment scores, using box plots or histograms.

Feature Engineering and Vectorization

- **Feature Extraction:** Convert text into numerical features using techniques like Bag of Words, TF-IDF, or word embeddings.

- **Dimensionality Reduction:** Use PCA or t-SNE to visualize feedback clusters in vector space.

Insights and Actionable Recommendations

- **Summarize Findings:** Highlight key patterns, recurring issues, and positive aspects.

- **Prioritize Actions:** Use insights to drive improvements in products, services, or customer support.

- **Monitor Trends:** Set up ongoing monitoring to track changes in customer sentiment over time.

Libraries that can be used

- **NLTK (Natural Language Toolkit):** A comprehensive library for text processing, offering tokenization, stopword removal, stemming, lemmatization, and basic sentiment analysis. NLTK is ideal for educational and prototyping purposes, although less optimized for large-scale applications.

- **spaCy:** A high-performance NLP library designed for practical use, supporting advanced features such as part-of-speech tagging, dependency parsing, and named entity recognition. spaCy is well-suited for large datasets and can be combined with sentiment analysis tools for a robust pipeline.

Feature	**NLTK**	**spaCy**
Main Focus	Education, research, prototyping	Production, efficiency, deployment
Performance	Slower, less optimized	Fast, highly optimized
Ease of Use	Flexible, but more complex	Simple, consistent API

Tokenization	Yes	Yes
POS Tagging	Yes	Yes
Named Entity Recognition (NER)	Yes (requires more code)	Yes (built-in, easy to use)
Sentiment Analysis	Yes (VADER, TextBlob)	Yes (requires integration/extension)

Sentiment Analysis Techniques

A. VADER (Valence Aware Dictionary and sEntiment Reasoner)

- VADER is a rule/lexicon-based sentiment analyzer well-suited for social media and short text.

- It provides a dictionary of scores for positive, negative, and neutral sentiment, along with a compound score ranging from -1 (most negative) to +1 (most positive).

- Implementation involved using NLTK's *SentimentIntensityAnalyzer* to assign sentiment scores to each text entry.

- **Findings:** VADER effectively captures overall sentiment trends, especially for clear-cut positive or negative reviews. However, it may not fully capture nuanced or context-dependent sentiments.

B. N-gram Analysis

- **Unigram and Bigram Frequency:** Extract and analyze the most common unigrams and bigrams to identify frequent words and phrases associated with different sentiment classes.

- N-gram analysis helps reveal patterns, such as common expressions in positive or negative reviews, and informs feature engineering for subsequent modeling.

- **Insights:** Positive reviews often contain words like "excellent," "love," and "great," while negative reviews feature terms like "disappointed," "poor," and "waste."