AI-Powered Smart Ticketing: Customer Complaints Classification

CAP5610 ML – Kalp Devangbhai Thakkar – 5548300 – ka765424@ucf.edu

GOAL

Automate the classification of customer complaints in the financial sector using advanced Natural Language Processing techniques. By leveraging machine learning models and topic modeling, the project aims to streamline complaint resolution and enhance operational efficiency in handling customer support tickets.

CONTENTS

CONTENTS

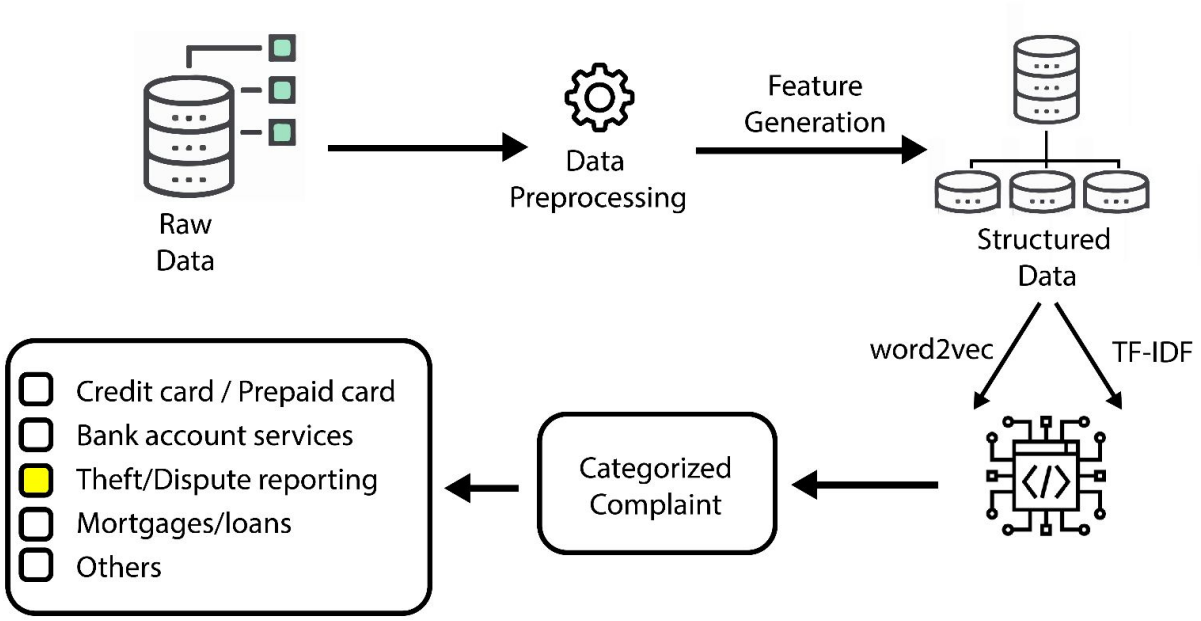
BACKGROUND AND OBJECTIVES

1. Rising customer complaints in financial institutions.
2. Manual classification leading to inefficiencies.
3. Objectives:
    - Automate complaint classification.
    - Improve resolution speed.
    - Enhance customer satisfaction.
    - Implement AI-driven support systems.

BACKGROUND AND OBJECTIVES

4. Utilize NLP and machine learning techniques.

5. Streamline ticketing processes for better service delivery.



ARCHITECTURE

WORKING PIPELINE

The pipeline involves preprocessing unstructured customer complaints, including data cleaning, lemmatization, and part-of-speech tagging. Next, Non-Negative Matrix Factorization (NMF) performs topic modeling to categorize complaints. Supervised models such as logistic regression, decision tree, and random forest are trained and evaluated to classify new complaints accurately, facilitating efficient resolution.

DATASET

DATA FORMAT

- **JSON format** provides flexibility and compatibility for data processing.

COMPLAINT COUNT

- A substantial dataset comprising **78,313 customer complaints** ensures comprehensive model training.

FEATURE DIVERSITY

- With **22 distinct features**, including text and categorical variables, the dataset offers a holistic view of customer feedback.


```
[{"_index": "complaint-public-v2",
"_type": "complaint", "_id":
"3211475", "_score": 0.0, "_source":
{"tags": null, "zip_code": "90301",
"complaint_id": "3211475", "issue":
"Attempts to collect debt not owed",
"date_received":
"2019-04-13T12:00:00-05:00",
"state": "CA", "consumer_disputed":
"N/A", "product": "Debt collection",
"company_response": "Closed with
explanation", "company": "JPMORGAN
CHASE & CO.", "submitted_via":
"Web", "date_sent_to_company":
"2019-04-13T12:00:00-05:00",
"company_public_response": null,
"sub_product": "Credit card debt",
"timely": "Yes",
"complaint_what_happened": "",
"sub_issue": "Debt is not yours",
"consumer_consent_provided":
"Consent not provided"}},{"_index":
"complaint-public-v2", "_type":
"complaint", "_id": "3229299",
" score": 0.0, " source": {"tags":
```

DATA LOADING

Leveraged the powerful pandas library to transform the JSON data into a structured DataFrame.

Organized the dataset into a structured format, facilitating preprocessing and exploratory analysis.

DATA PREPROCESSING

Column Renaming: Renamed columns for clarity and selected relevant features essential for project objectives.

Data Cleaning: Handled missing or irrelevant data to enhance dataset quality and suitability for analysis.

Noise Reduction: Identified and filtered out blank complaints to improve dataset quality and reduce noise levels.

TEXT PREPROCESSING

- Prepare textual data for exploratory analysis and topic modeling, laying a foundation for extracting meaningful insights and classifying complaints effectively.

TEXT STANDARDIZATION

All text converted to lowercase to ensure uniformity across the dataset, minimizing discrepancies due to letter casing variations.

ARTIFACT REMOVAL

Extraneous elements such as text within curly braces, line breaks, and numerical values systematically eliminated to streamline textual content.

PUNCTUATION REMOVAL

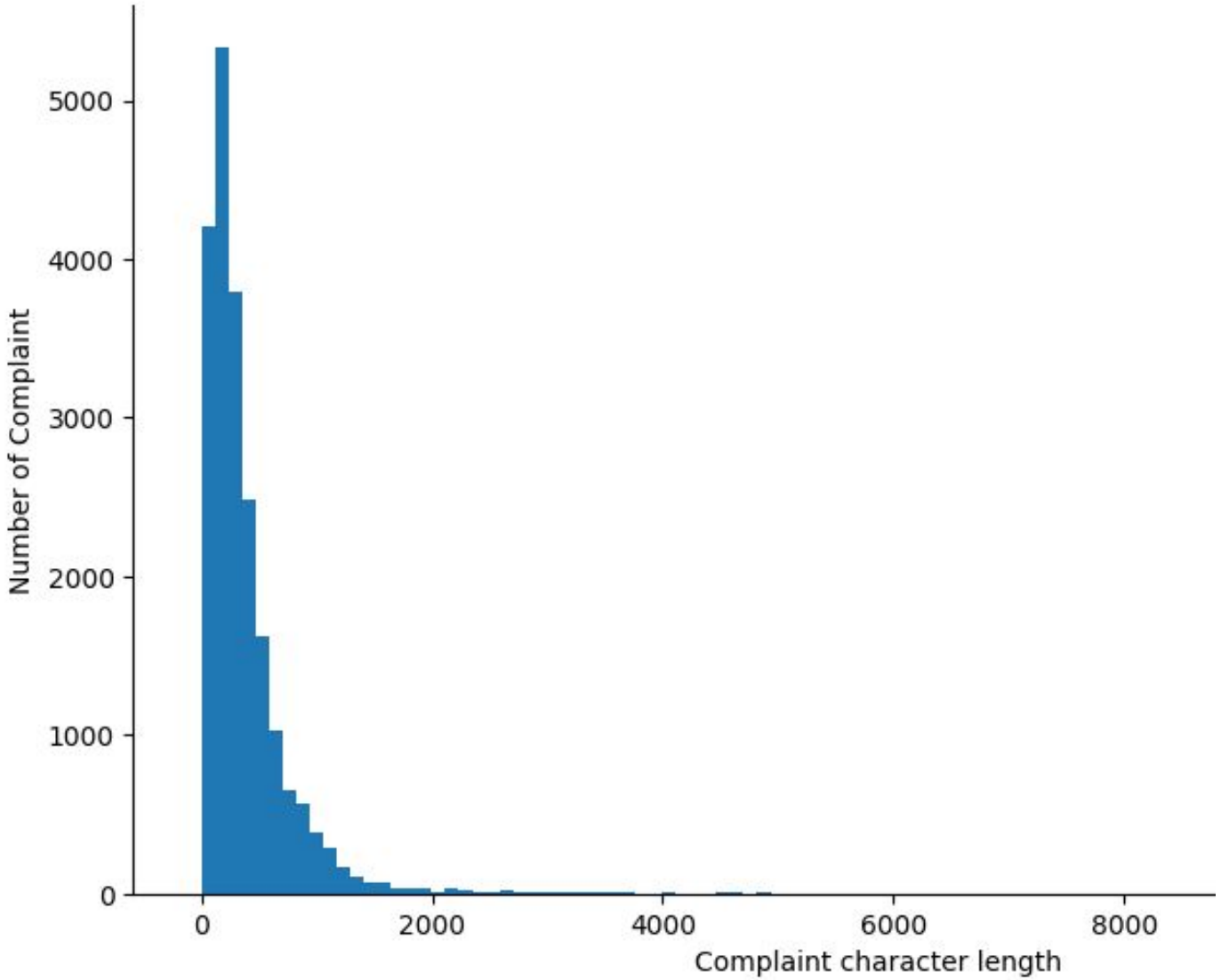
Punctuation systematically eliminated to enhance readability and facilitate subsequent analysis.

LEMMATIZATION

Leveraged NLTK library for lemmatization to reduce words to their base or root form, aiding in standardizing variations.

POS TAGGING

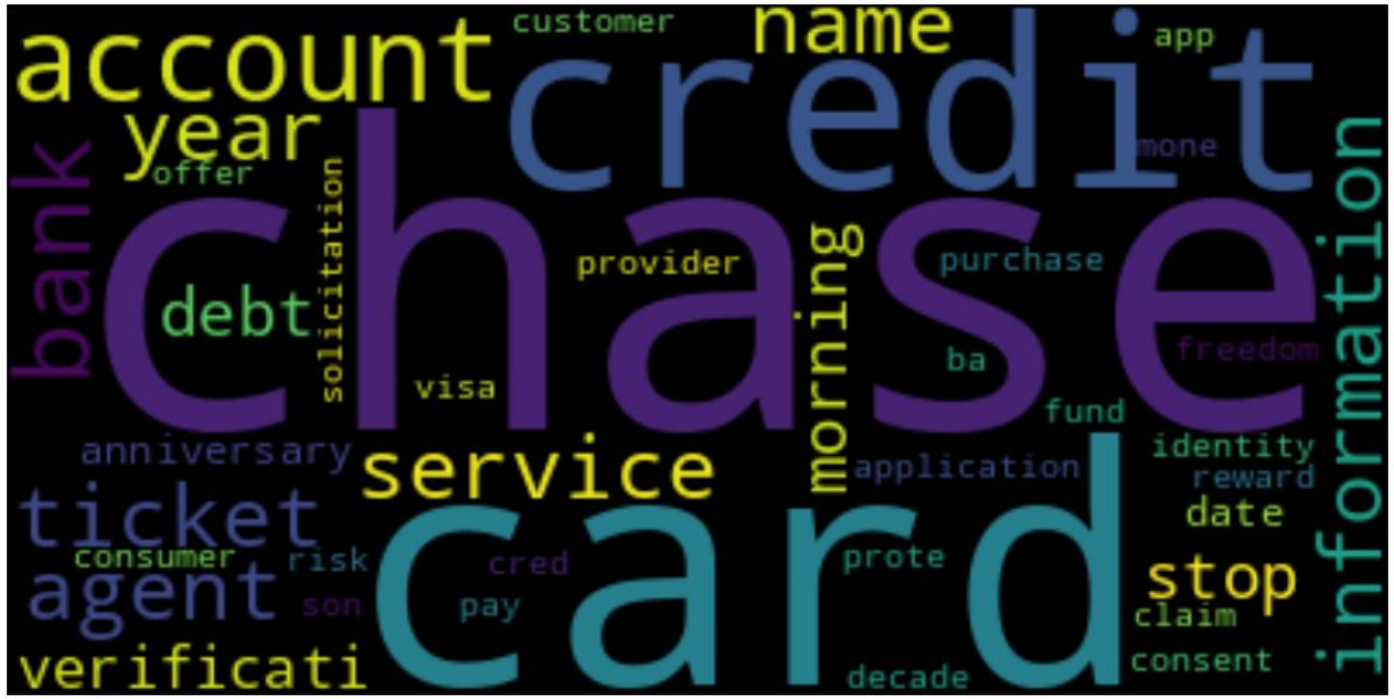
Extracted part-of-speech (POS) tags to discern the grammatical category of each word, with emphasis on retaining nouns (NN) for identifying core subject matter.



EXPLORATORY DATA ANALYSIS (EDA)

Distribution Analysis

Thorough examination of complaint lengths' distribution through histograms to identify significant variations and trends, offering insights into customer feedback nature.



EXPLORATORY DATA ANALYSIS (EDA)

Word Cloud Visualization

Utilized word cloud visualization to depict the top 40 words by frequency, providing a concise overview of prevalent themes and customer concerns.

EXPLORATORY DATA ANALYSIS (EDA)

- Explored unigrams, bigrams, and trigrams frequency to uncover linguistic patterns and recurring phrases, enhancing understanding of customer feedback.

Top 10 unigrams

```
[('chase', 54491),
 ('account', 47556),
 ('credit', 32787),
 ('card', 30436),
 ('bank', 21568),
 ('payment', 21367),
 ('time', 16311),
 ('day', 13322),
 ('charge', 12082),
 ('money', 12001)]
```

Top 10 bigrams

```
[('credit card', 12931),
 ('credit report', 3892),
 ('account chase', 3452),
 ('chase account', 3149),
 ('chase credit', 3097),
 ('customer service', 3082),
 ('account account', 3041),
 ('bank account', 2639),
 ('chase bank', 2244),
 ('debit card', 2067)]
```

Top 10 trigrams

```
[('chase credit card', 2027),
 ('credit card account', 1156),
 ('credit card company', 1005),
 ('credit card chase', 846),
 ('credit card credit', 589),
 ('inquiry credit report', 560),
 ('account credit card', 488),
 ('card credit card', 483),
 ('chase checking account', 394),
 ('credit report credit', 389)]
```

FEATURE EXTRACTION

- **Feature Extraction Objective:** Convert raw textual data into a numerical representation suitable for machine learning using TF-IDF technique.
- **TF-IDF Methodology:** Employed Term Frequency-Inverse Document Frequency (TF-IDF) to evaluate word importance relative to the document corpus.
- **TfidfVectorizer Implementation:** Utilized sklearn's TfidfVectorizer to initialize TF-IDF vectorizer with specified parameters.

FEATURE EXTRACTION

- **Parameter Optimization:** Set max_df and min_df parameters to 0.95 and 2, respectively, to filter out terms based on document frequency thresholds.

  max_df is set to 0.95, indicating the removal of terms that appear in more than 95% of the complaints. This helps eliminate corpus-specific stop words that may not provide valuable information for classification.

  min_df is set to 2, ensuring the removal of terms that appear in less than 2 complaints. This helps filter out infrequent terms that may not contribute significantly to the classification process.

FEATURE EXTRACTION

- **Document-Term Matrix Creation:** Created Document-Term Matrix (DTM) by applying fit_transform method to 'complaints' column in df_clean DataFrame.
- **DTM Representation:** DTM represents each complaint's TF-IDF scores for individual words, enabling accurate classification of complaints into relevant categories.

TOPIC MODELING

- Employ advanced NLP techniques, specifically Non-Negative Matrix Factorization (NMF), for uncovering latent patterns within customer complaints.
- NMF is an unsupervised technique that decomposes high-dimensional vectors into a lower-dimensional representation.
- The resulting vectors are non-negative, indicating that their coefficients are also non-negative.
- Implemented NMF to extract distinct clusters of related words representing specific topics across grievances.

ENABLING TOPIC-BASED CLASSIFICATION OF NEW COMPLAINTS THROUGH SUPERVISED LEARNING

- **Supervised Model Development:** Developed a supervised model to predict relevant topics for new complaints based on topics generated through topic modeling.
- **Topic Classification for New Complaints:** Utilized topics from topic modeling to classify new complaints effectively.
- **Conversion to Numerical Representations:** Converted topic names to numerical representations for compatibility with numpy arrays, essential for supervised learning techniques.
- **Seamless Integration:** Ensured smooth integration with supervised learning model for accurate classification of new complaints into respective topics.
- **Enhanced Efficiency:** Facilitated efficient processing and classification of new complaints through numerical representation of topics.

RESULTS - LOGISTIC REGRESSION

	precision	recall	f1-score	support
0	0.93	0.96	0.95	944
1	0.96	0.95	0.96	914
2	0.98	0.95	0.96	714
3	0.94	0.98	0.96	1123
4	0.97	0.90	0.93	520
accuracy			0.95	4215
macro avg	0.96	0.95	0.95	4215
weighted avg	0.95	0.95	0.95	4215

RESULTS - DECISION TREE

	precision	recall	f1-score	support
0	0.73	0.72	0.73	944
1	0.80	0.79	0.80	914
2	0.78	0.81	0.80	714
3	0.80	0.80	0.80	1123
4	0.69	0.69	0.69	520
accuracy			0.77	4215
macro avg	0.76	0.76	0.76	4215
weighted avg	0.77	0.77	0.77	4215

RESULTS - RANDOM FOREST CLASSIFIER

	precision	recall	f1-score	support
0	0.83	0.63	0.72	944
1	0.75	0.81	0.78	914
2	0.89	0.79	0.83	714
3	0.60	0.97	0.74	1123
4	1.00	0.12	0.21	520
accuracy			0.72	4215
macro avg	0.81	0.66	0.66	4215
weighted avg	0.78	0.72	0.69	4215

RESULTS - GAUSSIAN NAIVE BAYES

	precision	recall	f1-score	support
0	0.48	0.35	0.40	944
1	0.35	0.27	0.30	914
2	0.52	0.50	0.51	714
3	0.46	0.30	0.37	1123
4	0.18	0.51	0.27	520
accuracy			0.36	4215
macro avg	0.40	0.38	0.37	4215
weighted avg	0.42	0.36	0.37	4215

CONCLUSION

Logistic regression excels among machine learning models in classifying customer complaints, achieving an impressive **95% accuracy**

CHALLENGES AND LIMITATIONS

- Data Quality and Completeness
- Complexity of Natural Language
- Optimal Model Selection
- Interpretable
- Topic Modeling
- Scalability and Efficiency
- Generalization Across Categories
- Model Evaluation Metrics

REFERENCES

1] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media. [Online]. Available: http://www.nltk.org/book/
[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,. . . Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. [Online]. Available: https://scikit-learn.org/stable/documentation.html
[3] Explosion AI. (2021). SpaCy 3.1 Documentation. [Online]. Available: https://spacy.io/api
[4] A collection of Python code examples for WordCloud visualization. [Online].Available: https://github.com/amueller/word_cloud
[5] Plotly Technologies Inc. (2021). Plotly Python Open Source Graphing Library. [On-line]. Available: https://plotly.com/python/
[6] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negativematrix factorization. Nature, 401(6755), 788-791.
[7] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual Analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.
[8] Verhoef, P. C., Lemon, K. N., Parasuraman, A., Roggeveen, A., Tsiros, M., & Schlesinger, L. A. (2009). Customer experience creation: Determinants, dynamics and management strategies. Journal of Retailing, 85(1), 31-41.

THANK YOU

Kalp Devangbhai Thakkar
UCFID: 5548400