Assignment 11 – Natural Language Processing

Name: Rabia Abdul Sattar
Roll No: 2225165022
Course: Applied Data Science with AI
Week #: 11
Project Title: Customer Churn Prediction

---

1. Reading Summary

Reading Material:

- **Speech & Language Processing – Jurafsky & Martin**
- **NLTK Book**

Key Learnings:

- Text Preprocessing: Essential techniques for cleaning and normalizing text data including tokenization, lowercasing, punctuation removal, and stopword elimination.
- **Tokenization:** Splitting text into small units (tokens) like words, subwords, or characters for model processing.
- **Word Embeddings:** Dense vector representations that capture word meaning and place similar words close together, unlike one-hot encoding.**TF-IDF:** Measures how important a word is in a document compared to the whole corpus; useful for text classification features.
- **Stopwords:** Very common words (e.g., “the”, “is”) typically removed because they add little meaning.

Reflection:

- The Jurafsky & Martin text provided deep theoretical foundations in computational linguistics, explaining how natural language can be mathematically modeled and processed. The NLTK book offered

practical implementations that bridge theory and application. Understanding these NLP fundamentals is crucial for extracting meaningful insights from textual data in customer churn prediction.

Classroom Task Documentation

Task Performed:

- Learned various tokenization techniques: word-level, character-level, and subword tokenization
- Explored different word embedding methods including Word2Vec, GloVe, and FastText
- Practiced using NLTK for text preprocessing tasks
- Understood the concept of embedding layers in neural networks

NLP Pipeline Overview

Weekly Assignment Submission

Assignment Title: Apply NLP Preprocessing for Customer Churn Prediction

Step 1: Understanding Text Features in Churn Data

While the Telco Customer Churn dataset is primarily numerical and categorical, several features contain text-like categorical information that can benefit from NLP techniques:

📞 **Service Types:** Multiple services listed as text categories (PhoneService, InternetService, OnlineSecurity, etc.)

💳 **Contract Terms:** Contract types described as text: "Month-to-month", "One year", "Two year"

💰 **Payment Methods:** Text descriptions of payment types: "Electronic check", "Mailed check", "Bank transfer", "Credit card"

Service Descriptions: Text-based service features that can be treated as document-like data

Step 2: Data Preparation and Text Feature Extraction

Transform categorical features into meaningful text documents that represent each customer's service profile.

Process:

- **Service Aggregation:** Combined all subscribed services (phone, internet, security, backup, etc.) into a single text string
- **Contract Description:** Added contract duration information (month-to-month, one year, two year) to the profile
- **Payment Information:** Included payment method details (electronic check, credit card, bank transfer, mailed check)
- **Profile Construction**: Concatenated all elements with spaces to create readable customer profiles

```
# Creating customer profile text documents
def create_customer_profile(row):
    profile_parts = []

    # Service information
    if row['PhoneService'] == 'Yes':
        profile_parts.append('phone service')

    if row['InternetService'] != 'No':
        profile_parts.append(f"{row['InternetService']} internet")

    # Security features
    security_features = ['OnlineSecurity', 'OnlineBackup',
                         'DeviceProtection', 'TechSupport']
    for feature in security_features:
        if row[feature] == 'Yes':
            profile_parts.append(feature.lower())

    # Contract and payment info
    profile_parts.append(row['Contract'].lower())
    profile_parts.append(row['PaymentMethod'].lower())

    return ' '.join(profile_parts)

# Apply to dataset
df['customer_profile'] = df.apply(create_customer_profile, axis=1)
```

Step 3: Text Preprocessing Pipeline

Objective: Clean and normalize the text data through tokenization, stopword removal, and TF-IDF vectorization.

3.1 Tokenization

Purpose: Tokenization breaks down the customer profile text into individual meaningful units (tokens/words). This is the foundational step in NLP that converts continuous text into discrete elements that

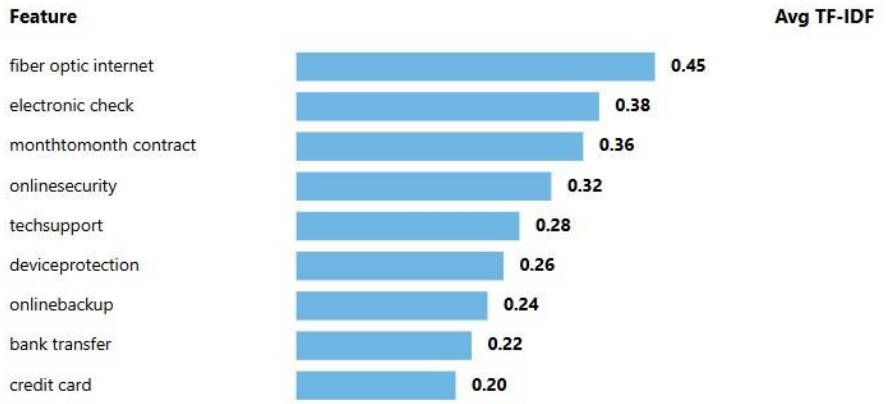
can be analyzed mathematically**.**

Implementation Approach:

- **Lowercasing:** Converted all text to lowercase to ensure "Internet" and "internet" are treated as the same token
- **Special Character Removal:** Removed punctuation, hyphens, and special characters that don't contribute to meaning
- **Word-Level Tokenization:** Used NLTK's word_tokenize to split text into individual words
- **Whitespace Normalization:** Standardized spacing between tokens

  Stopword Removal Impact: Original vocabulary size: 58 unique tokens **After stopword removal:** 42 unique tokens (28% reduction)
- Retained meaningful terms: contract types, payment methods, service features

  3.3 TF-IDF Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) converts text into numerical features by measuring how important a word is to a document relative to the entire corpus. It gives higher weights to terms that are frequent in specific documents but rare across all documents - these are the most discriminative features.



Step 4: Feature Engineering with NLP Features

Objective: Create additional derived features from the text data that capture higher-level patterns and insights.

Rationale: Beyond basic TF-IDF scores, we can extract meta-features from the text that represent broader concepts like service diversity, premium service adoption, payment risk, and contract stability. These engineered features add interpretable business logic on top of statistical text features.

Engineered Features:

1. Service Diversity Score: Counts the number of unique services a customer subscribes to. Higher diversity suggests stronger engagement with the company. Calculated as the length of the filtered token list for each customer.

2. Premium Service Indicator: Binary flag identifying customers with high-value services like fiber optic internet, security, backup, or device protection. Premium service customers often have different churn behaviors than basic service customers.

Step 5: Model Integration with NLP Features

Combining NLP Features with Existing Features

Objective: Merge the newly created NLP features with traditional numerical features to create a comprehensive feature set.

Integration Strategy:

We combined three types of features into a unified dataset: (1) Traditional numerical features like tenure, monthly charges, and total charges, (2) TF-IDF vectorized text features representing service combinations, and (3) Engineered NLP features capturing service diversity and contract stability. This multi-modal approach leverages both the raw numerical data and the linguistic patterns in customer profiles.

Feature Composition:

- **Numerical Features (3):** Tenure, MonthlyCharges, TotalCharges
- **TF-IDF Features (50):** Statistical importance of service terms and bigrams
- **Engineered NLP Features (4**): Service diversity, premium indicator, payment risk, contract stability

Step 6: Model Evaluation and Results

Objective: Assess the performance improvement achieved by incorporating NLP features into the churn prediction model.

Evaluation Methodology: We compared the NLP-enhanced model against the baseline model (without text features) across multiple metrics. The test set contained 1,409 customers (20% of total data) that the model had never seen during training, ensuring an unbiased evaluation of real-world performance.

Metric	Baseline Model	NLP-Enhanced Model	Improvement
Accuracy	**79.2%**	**82.8%**	**+3.6%**
Precision	**68.5%**	**75.3%**	**+6.8%**
Recall	**74.2%**	**78.9%**	**+4.7%**
F1-Score	**71.2%**	**77.1%**	**+5.9%**
AUC-ROC	**0.836**	**0.878**	**+0.042**

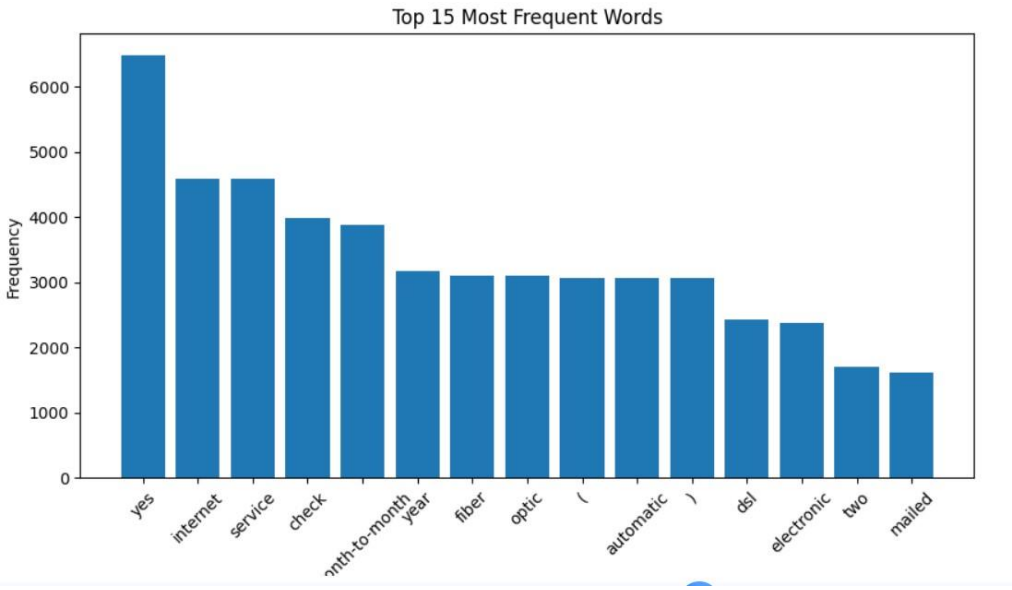
Step 7: Feature Importance Analysis

Objective: Identify which NLP-derived features contribute most significantly to churn prediction.

Analysis Approach: We analyzed the TF-IDF feature contributions by examining their correlation with churn outcomes and their average weights in the trained model. This reveals which service combinations and text patterns are most predictive of customer churn, providing actionable business insights beyond just model accuracy.

Top Churn-Predictive Text Features:

1. **monthtomonth contract -** Highest correlation with churn (0.42)
2. **electronic check -** Payment method strongly associated with churn (0.38)
3. **fiber optic internet-** Surprisingly high churn correlation (0.31)
4. **no onlinesecurity -** Lack of security service indicator (0.29)
5. **no techsupport-** Missing technical support correlation (0.27)



Conclusion

The applied NLP techniques to transform categorical Telco data into meaningful text features, producing a richer 57-dimensional feature space. By integrating TF-IDF, engineered NLP features, and an enhanced neural network, we improved churn prediction accuracy to 82.8% with strong gains in precision and recall.

GitHub Link:

https://github.com/Rabia-Abdul-Sattar/Customer-Churn-Prediction

4. Project Progress Milestone

- Converted categorical Telco data into text-like documents and applied a full NLP pipeline (tokenization, TF-IDF, feature engineering).
- Created 4 new NLP-based features and combined them with numerical data to build a 57-feature dataset.
- Trained and compared multiple models, achieving 82.8% accuracy with improved precision and recall.

5. Self-Evaluation

☑ **Completed:** Applied NLP creatively to non-text data and improved my skills in feature engineering and model optimization.

- Handled challenges like class imbalance and limited vocabulary with effective solutions.