2403A52006

MD.Mustafa

Loading the Dataset

```python
import pandas as pd

# Load dataset
df = pd.read_csv("Tweets.csv")

# Check basic info
df.head()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 14640,\n  \"fields\":
[\n    {\n      \"column\": \"tweet_id\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 779111158481836,\n
\"min\": 567588278875213824,\n        \"max\": 570310600460525568,\n
\"num_unique_values\": 14485,\n        \"samples\": [\n
567917894144770049,\n          567813976492417024,\n
569243676594941953\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"airline_sentiment\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 3,\n        \"samples\":
[\n          \"neutral\",\n          \"positive\",\n
\"negative\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"airline_sentiment_confidence\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0.1628299590986659,\n
\"min\": 0.335,\n        \"max\": 1.0,\n        \"num_unique_values\":
1023,\n        \"samples\": [\n          0.6723,\n          0.3551,\n
0.6498\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"negativereason\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 10,\n
\"samples\": [\n          \"Damaged Luggage\",\n          \"Can't
Tell\",\n          \"Lost Luggage\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"negativereason_confidence\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
0.3304397596377413,\n        \"min\": 0.0,\n        \"max\": 1.0,\n
\"num_unique_values\": 1410,\n        \"samples\": [\n
0.6677,\n          0.6622,\n          0.6905\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
6,\n        \"samples\": [\n          \"Virgin America\",\n
\"United\",\n          \"American\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline_sentiment_gold\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n
\"negative\",\n          \"neutral\",\n          \"positive\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"name\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\":
7701,\n        \"samples\": [\n          \"smckenna719\",\n
\"thisAnneM\",\n          \"jmspool\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"negativereason_gold\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 13,\n        \"samples\": [\n
\"Customer Service Issue\\nLost Luggage\",\n          \"Late Flight\\
nCancelled Flight\",\n          \"Late Flight\\nFlight Attendant
Complaints\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"retweet_count\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 0,\n        \"min\": 0,\n
\"max\": 44,\n        \"num_unique_values\": 18,\n        \"samples\":
[\n          0,\n          1,\n          6\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"text\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 14427,\n
\"samples\": [\n          \"@JetBlue so technically I could drive to
JFK now and put in. Request for tomorrow's flight?\",\n
\"@united why I won't check my carry on. Watched a handler throw this
bag -- miss the conveyer belt -- sat there 10 min
http://t.co/lyoocx5mSH\",\n          \"@SouthwestAir you guys are so
clever \\ud83d\\ude03 http://t.co/qn5odUGFqK\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"tweet_coord\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 832,\n        \"samples\": [\n
\"[40.04915451, -75.10364317]\",\n          \"[32.97609561, -
96.53349238]\",\n          \"[26.37852293, -81.78472152]\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"tweet_created\",\n      \"properties\": {\n        \"dtype\":
\"object\",\n        \"num_unique_values\": 14247,\n
\"samples\": [\n          \"2015-02-23 07:40:55 -0800\",\n
\"2015-02-21 16:20:09 -0800\",\n          \"2015-02-21 21:33:21 -
0800\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"tweet_location\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 3081,\n
\"samples\": [\n          \"Oakland, California\",\n
\"Beverly Hills, CA\",\n          \"Austin, TX/NY, NY\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"user_timezone\",\n      
\"properties\": {\n        \"dtype\": \"category\",\n        
\"num_unique_values\": 85,\n        \"samples\": [\n          
\"Helsinki\",\n          \"Eastern Time (US & Canada)\",\n          
\"America/Detroit\"\n        ],\n        \"semantic_type\": \"\",\n        
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df"}
```

Dataset Exploration

```
df.shape
df.columns
```

```
Index(['tweet_id', 'airline_sentiment',
'airline_sentiment_confidence',
       'negativereason', 'negativereason_confidence', 'airline',
       'airline_sentiment_gold', 'name', 'negativereason_gold',
       'retweet_count', 'text', 'tweet_coord', 'tweet_created',
       'tweet_location', 'user_timezone'],
      dtype='object')
```

Selecting Relevant Columns

```
df = df[['text', 'airline_sentiment']]
df.head()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 14640,\n  \"fields\":
[\n    {\n      \"column\": \"text\",\n      \"properties\": {\n        
\"dtype\": \"string\",\n        \"num_unique_values\": 14427,\n        
\"samples\": [\n          \"@JetBlue so technically I could drive to 
JFK now and put in. Request for tomorrow's flight?\",\n          
\"@united why I won't check my carry on. Watched a handler throw this 
bag -- miss the conveyer belt -- sat there 10 min 
http://t.co/lyoocx5mSH\",\n          \"@SouthwestAir you guys are so 
clever \\ud83d\\ude03 http://t.co/qn5odUGFqK\"\n        ],\n        
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline_sentiment\",\n      
\"properties\": {\n        \"dtype\": \"category\",\n        
\"num_unique_values\": 3,\n        \"samples\": [\n          
\"neutral\",\n          \"positive\",\n          \"negative\"\n        
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n      
}\n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

Install & Import Required Libraries

```
import re
import nltk
import spacy
import matplotlib.pyplot as plt
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
from wordcloud import WordCloud

nltk.download('stopwords')
nltk.download('punkt')

from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
```

```
!python -m spacy download en_core_web_sm
```

```
Collecting en-core-web-sm==3.8.0
  Downloading
https://github.com/explosion/spacy-models/releases/download/en_core_we
b_sm-3.8.0/en_core_web_sm-3.8.0-py3-none-any.whl (12.8 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 12.8/12.8 MB 105.0 MB/s eta
0:00:00
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_sm')
⚠ Restart to reload dependencies
If you are in a Jupyter or Colab notebook, you may need to restart
Python in
order to load all the package's dependencies. You can do this by
selecting the
'Restart kernel' or 'Restart runtime' option.
```

```
nlp = spacy.load("en_core_web_sm")
```

Text Preprocessing

```
def clean_tweet(text):
    text = re.sub(r"http\S+|www\S+", "", text)   # URLs
    text = re.sub(r"@\w+", "", text)             # mentions
    text = re.sub(r"#\w+", "", text)             # hashtags
    text = re.sub(r"[^a-zA-Z\s]", "", text)      # special chars
    text = text.lower()
    return text

df['clean_text'] = df['text'].apply(clean_tweet)
df.head()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 14640,\n  \"fields\":
[\n    {\n      \"column\": \"text\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 14427,\n
\"samples\": [\n          \"@JetBlue so technically I could drive to
JFK now and put in. Request for tomorrow's flight?\",\n
\"@united why I won't check my carry on. Watched a handler throw this
bag -- miss the conveyer belt -- sat there 10 min
http://t.co/lyoocx5mSH\",\n          \"@SouthwestAir you guys are so
clever \\ud83d\\ude03 http://t.co/qn5odUGFqK\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline_sentiment\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n
\"neutral\",\n          \"positive\",\n          \"negative\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"clean_text\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 14203,\n        \"samples\": [\n          \"
bag is supposedly here in boston\",\n          \" thanks we would like
an apology and a full refund please confirm when we will receive\",\n
\" i would like to thank the customer service team for their response
to my cancelled flightled flight but just offering to cont\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df"}
```

Tokenization and Stopword Removal

```python
import nltk
nltk.download('punkt_tab')

stop_words = set(stopwords.words('english'))

def tokenize(text):
    tokens = nltk.word_tokenize(text)
    tokens = [word for word in tokens if word not in stop_words]
    return tokens

df['tokens'] = df['clean_text'].apply(tokenize)
df.head()
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 14640,\n  \"fields\":
[\n    {\n      \"column\": \"text\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 14427,\n
\"samples\": [\n          \"@JetBlue so technically I could drive to
JFK now and put in. Request for tomorrow's flight?\",\n
\"@united why I won't check my carry on. Watched a handler throw this
bag -- miss the conveyer belt -- sat there 10 min
http://t.co/lyoocx5mSH\",\n          \"@SouthwestAir you guys are so
clever \\ud83d\\ude03 http://t.co/qn5odUGFqK\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"airline_sentiment\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n
\"neutral\",\n          \"positive\",\n          \"negative\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"clean_text\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 14203,\n        \"samples\": [\n          \"
bag is supposedly here in boston\",\n          \" thanks we would like
an apology and a full refund please confirm when we will receive\",\n
\" i would like to thank the customer service team for their response
to my cancelled flightled flight but just offering to cont\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"tokens\",\n      \"properties\": {\n        \"dtype\": \"object\",\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

Filtering Negative Sentiment Tweets

```
negative_df = df[df['airline_sentiment'] == 'negative']
negative_df.shape
```

```
(9178, 4)
```

```
negative_text = negative_df['tokens'].apply(lambda x: " ".join(x))
```

TF-IDF Feature Extraction

```
tfidf = TfidfVectorizer(max_features=20)
tfidf_matrix = tfidf.fit_transform(negative_text)

feature_names = tfidf.get_feature_names_out()
scores = tfidf_matrix.sum(axis=0).A1
```

TF-IDF Score Analysis

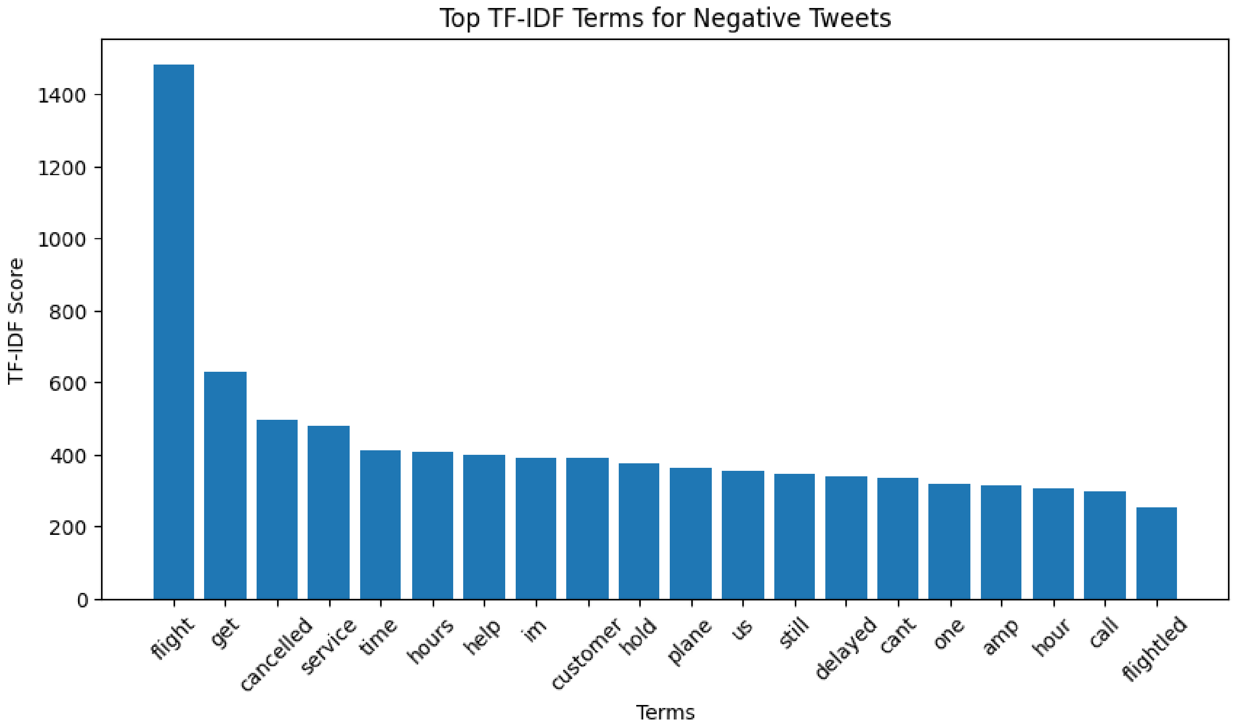
```
tfidf_df = pd.DataFrame({
    'Term': feature_names,
    'TF-IDF Score': scores
}).sort_values(by='TF-IDF Score', ascending=False)

tfidf_df
```

```
{"summary":"{\n  \"name\": \"tfidf_df\",\n  \"rows\": 20,\n
\"fields\": [\n    {\n      \"column\": \"Term\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 20,\n        \"samples\": [\n
\"flight\",\n          \"hour\",\n          \"one\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"TF-IDF Score\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
```

```
259.52230193290166,\n        \"min\": 255.40850214144848,\n
\"max\": 1480.9999997238308,\n        \"num_unique_values\": 20,\n
\"samples\": [\n          1480.9999997238308,\n
308.30843558068057,\n          317.4015525746276\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"tfidf_df"}
```

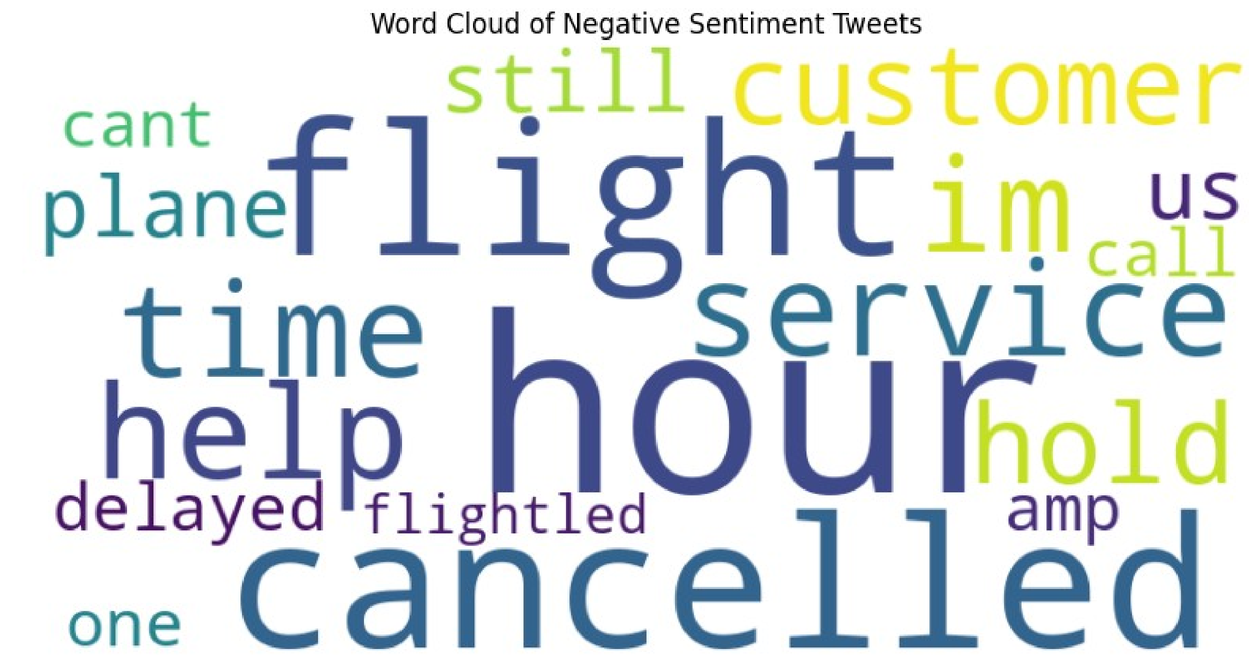
Visualization Using Bar Chart

```
plt.figure(figsize=(10,5))
plt.bar(tfidf_df['Term'], tfidf_df['TF-IDF Score'])
plt.xticks(rotation=45)
plt.title("Top TF-IDF Terms for Negative Tweets")
plt.xlabel("Terms")
plt.ylabel("TF-IDF Score")
plt.show()
```



Word Cloud Visualization

```
wordcloud = WordCloud(
    width=800,
    height=400,
    background_color='white'
).generate(" ".join(tfidf_df['Term']))

plt.figure(figsize=(10,5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title("Word Cloud of Negative Sentiment Tweets")
plt.show()
```



Discussion

In this experiment, Twitter airline sentiment data was analyzed using NLP preprocessing techniques such as text cleaning, tokenization, and stopword removal. TF-IDF was applied to negative sentiment tweets to identify the most important terms contributing to customer dissatisfaction.

The results show that words related to flight delays, cancellations, and customer service have higher TF-IDF scores, indicating common complaint themes. Bar chart and word cloud visualizations effectively highlight dominant negative sentiment vocabulary. Overall, TF-IDF proves to be a useful technique for extracting meaningful insights from social media text data.