Logo, company name

Description automatically generated

**NLP Analysis of Customer Support Chat Data**

**Natural Language Processing**

**PUSL3189**

**Lanka Pathmakumara**

**10899186**

**Contents**

[**1.** **Introduction** 3](#_Toc182403164)

[**2.** **Dataset Overview** 3](#_Toc182403165)

[**3.** **Data Preprocessing** 3](#_Toc182403166)

[**3.1.** **Tokenization** 3](#_Toc182403167)

[**3.1.1.** **Sentence Tokenization** 3](#_Toc182403168)

[**3.1.2.** **Word Tokenization** 4](#_Toc182403169)

[**3.2.** **Stemming** 4](#_Toc182403170)

[**3.3.** **Lemmatization** 5](#_Toc182403171)

[**4.** **POS Tagging** 6](#_Toc182403172)

[**5.** **Named Entity Recognition (NER)** 7](#_Toc182403173)

[**6.** **Analysis & Insights** 8](#_Toc182403174)

[**6.1.** **Common Issues** 8](#_Toc182403175)

[**6.2.** **Topic Discovery** 8](#_Toc182403176)

[**6.3.** **Sentiment Insight** 9](#_Toc182403177)

[**7.** **Conclusion** 10](#_Toc182403178)

# **Introduction**

The company "QuickHelp" handles a large number of support inquiries based on various products, and this report provides an NLP-based analysis of customer support chat data for this company. The main objectives of this analysis are to find recurring themes, common customer problems and ways to improve chatbot responses. We examine text data using methods such as named entity recognition (NER), stacking, lemmatization, POS tagging, tokenization, and stacking to learn more about client query types and topic word usage trends. This work offers a holistic way to understand the most prevalent issues and sentiments for customer queries, password resets, order tracking and to find out where customers are facing common issues. With these findings, "QuickHelp" can increase the efficiency of its chatbot and streamline support services, ultimately improving customer satisfaction and reducing the need for human intervention in common queries.

# **Dataset Overview**

A chat log data set containing details of customer inquiries received by a support team was used for this investigation and is intended to assist in analyzing customer issues, finding common issues and gaining knowledge to improve the support experience.

customer\_id: A unique number assigned to identify each customer.

chat\_date: The time and date each chat started. It can be used as separate date and time fields or as a single format (YYYY-MM-DD HH:MM).

message\_text: The actual text of the customer's query.

# **Data Preprocessing**

Data preprocessing is a critical first step in converting unstructured text data into an analysis ready format. Lemmatization, stemming and tokenization are important methods for standardizing and simplifying text.  
  
Tokenization facilitates the analysis of specific syntactic components by breaking text into smaller units, usually words or phrases. Word suffixes are removed through stemming, creating a root form that is sometimes inaccurate but effective. This is further refined by lemmatization, which reduces words to their lexical base form while keeping the meaning intact. When combined, these techniques prepare data for more complex processes such as sentiment analysis and pattern recognition.

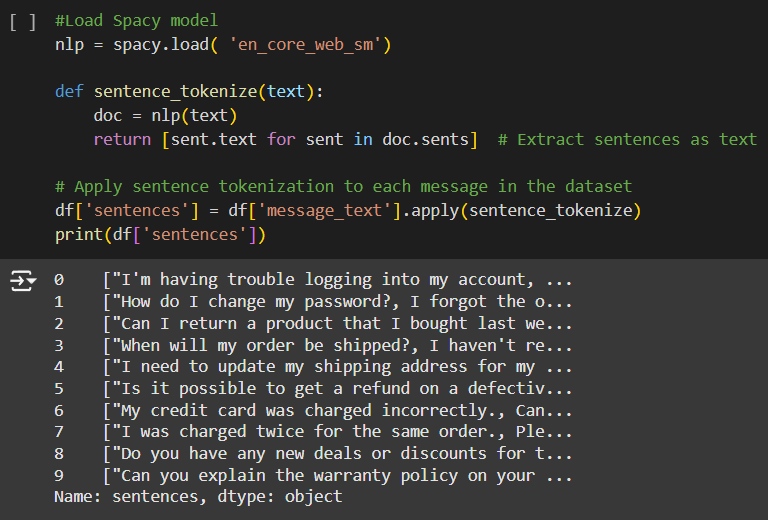
## **Tokenization**

The process of tokenizing text involves dividing it into discrete units that may be handled independently, usually words or sentences.

### **Sentence Tokenization**

When examining client communications at the sentence level, dividing the content into separate sentences, this phase is very useful because it enables us to identify specific problems or queries within each query.

The below code follows tokenizes every client message into sentences using the spaCy library:



### **Word Tokenization**

Separates text into words and eliminates whitespace and punctuation. Deeper examination of the precise words used is made possible by this degree of detail, which also helps to detect important phrases, themes, and sentiment in customer communications.

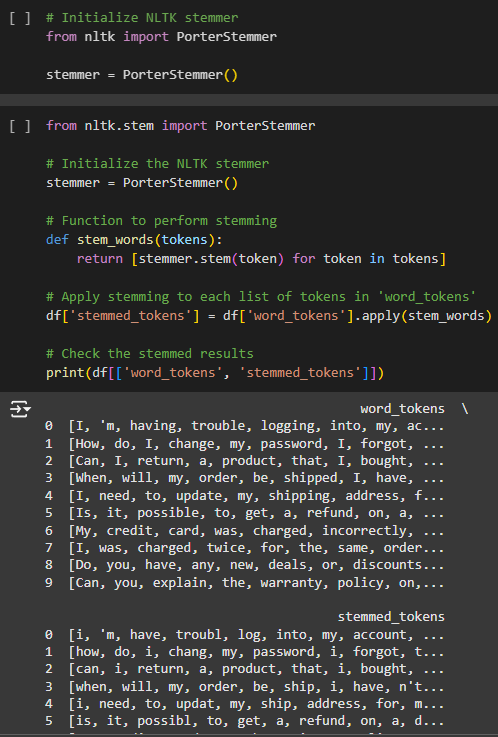
A screenshot of a computer screen

Description automatically generatedEach message is subjected to word tokenization by the following code, which removes whitespace and punctuation:

## **Stemming**

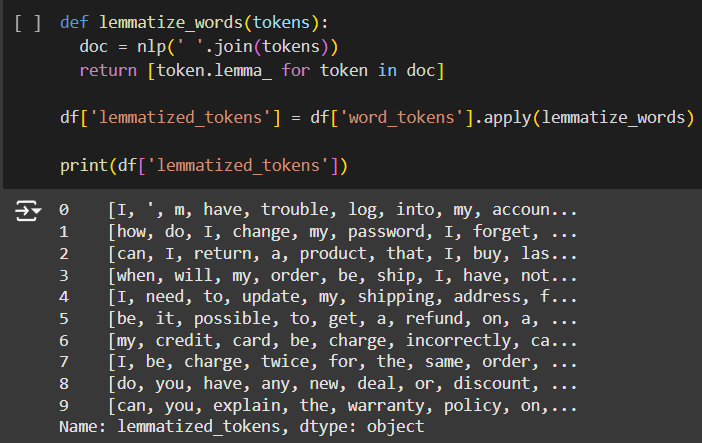
A text preprocessing method called stemming breaks words down to their most basic or root form. Stemming makes it easy to assess the fundamental meaning of terms in various forms by standardizing text and grouping comparable words by removing suffixes (e.g., "running" and "runner" are reduced to "run").

NLTK's PorterStemmer, a well-known stemmer that reduces words to their roots by eliminating frequent suffixes, is used in the following code:



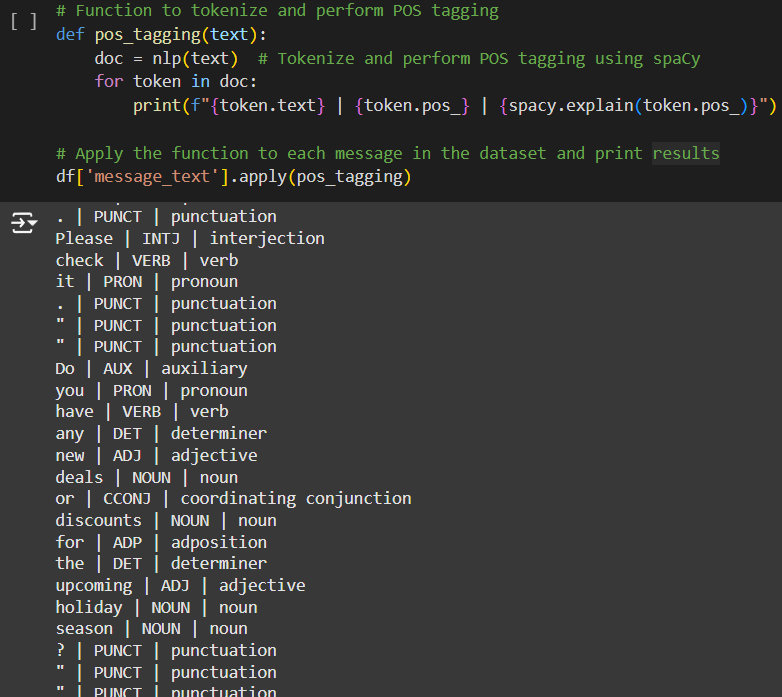
## **Lemmatization**

Another text normalization technique is lemmatization, which breaks down words into their basic or lexical forms, or "lemmas." Lemmatization, in contrast to transposition, takes into account a word's context and uses lexical information to ensure that the root form of a word is still a recognized word in the language. For example, it maintains grammatical correctness by changing "better" to "good" and "run" to "run."

The spaCy library is used for lemmatization in this code. Accurate lemmas are produced by using spaCy's language model to parse tokenized words:

# **POS Tagging**

Each text token is assigned a part of speech via the POS tagging method, which is useful for understanding language structure and can support various NLP tasks such as entity recognition, dependency parsing, and text classification.

The following code tokenizes and tags each word with its matching POS using spaCy. Every word is printed along with its POS tag and a brief explanation as it iterates over each message and applies POS tagging.

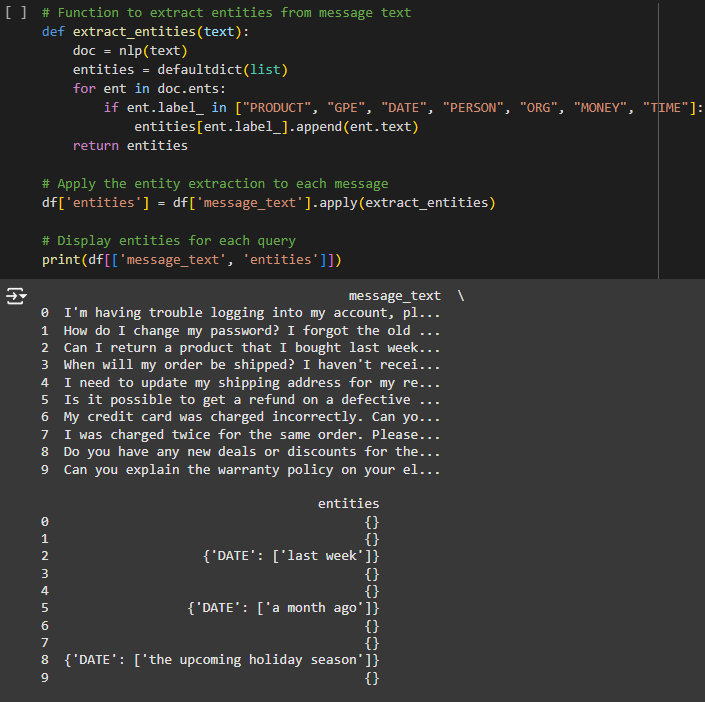
A screen shot of a computer

Description automatically generatedThe get\_pos\_counts method in the following code gathers the number of POS tags for every message, disregarding punctuation and spaces. It counts the instances of each POS tag across all tokens using the Counter class.

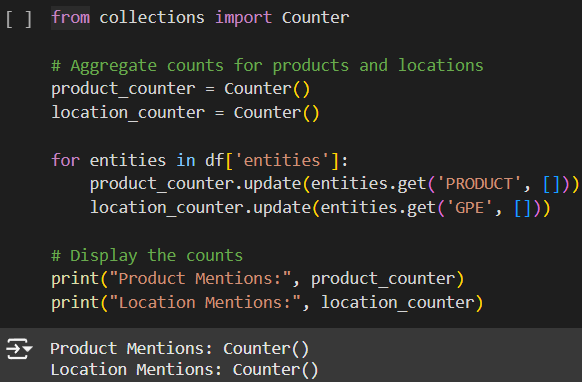
# **Named Entity Recognition (NER)**

The main objective is to use Named Entity Recognition (NER) to extract relevant entities from text messages such as product names, locations, dates and customer names, while also finding trends and patterns such as products or areas that may require further support.

We extract named entities (such as product names, places, dates, and customer names) from every communication using the spacy model. There are several types of entities, including PRODUCT, GPE (Geopolitical Entity), DATE, PERSON, ORG, and so on.



Next, we combine the product and location counts from all of the dataset's messages. We notice that the data for product and location mentions is not currently available since no product mentions were recovered (for example, because the messages did not contain particular product names) and the locations collected are empty.



# **Analysis & Insights**

## **Common Issues**

We tokenized the messages, filtered out stopwords, and counted the occurrences of each token in order to find commonly occurring words associated with complaints or frequently asked questions.

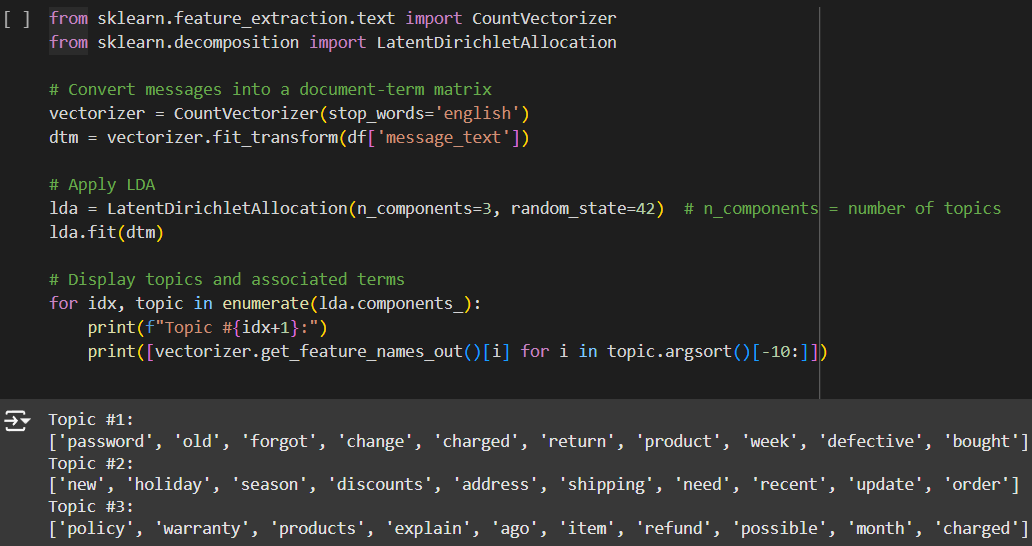
The terms "please" (two mentions), "bought" (two mentions), and "order" (two mentions) are the most frequently used complaint terms. "charged" is mentioned twice. The terms "trouble," "account," "help," and so forth are also frequently used. These frequently used terms suggest that a large number of client inquiries are to problems with logging in, orders, charges, and help requests, particularly in relation to orders and products.

A screen shot of a computer

Description automatically generated

## **Topic Discovery**

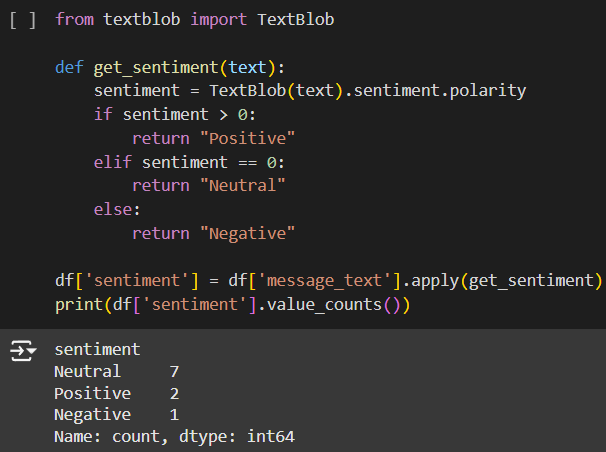
Three recurrent themes emerged from the consumer questions using the topic modeling technique known as Latent Dirichlet Allocation (LDA). The subjects were shown in the picture below:



These topics reflect common issues with order issues, refund policies, product warranties, and shipping status, helping to identify areas where customer support is most frequently needed.

## **Sentiment Insight**

Based on the polarity score, we classified each query as positive, neutral, or negative using sentiment analysis using TextBlob, a simple sentiment analysis tool.



Most consumer inquiries are neutral, meaning they are either generic questions or requests for information. While some of the queries are negative, suggesting discontent or problems that require attention, a tiny percentage are favorable, probably expressing client happiness.

# **Conclusion**

An overview of the key conclusions and findings follows.Examining the customer support chat data revealed a number of recurring themes and customer issues. Customers often complain about problems accessing accounts, changing passwords, returning faulty goods, getting shipment updates and requesting refunds. Several positive and negative attitudes were noted, however neutral query sentiment dominated the study. Three main topics emerged from topic modeling: product and order issues, shipping and address changes, and policy-related questions such as returns and warranties. By emphasizing entities such as items, places, and dates, Named Entity Identification (NER) also helped clarify where and when specific client problems occurred.

**GitHub Link -** [**https://github.com/HChandeepa/NLP\_Analysis\_of\_Customer\_Support\_Chat\_Data**](https://github.com/HChandeepa/NLP_Analysis_of_Customer_Support_Chat_Data)