# An Introduction to Natural Language Processing

Gurbuz Kutay Turkoglu

May 27, 2021

# Contents

**Abstract**

Natural Language Processing is a very broad field which can be used in different areas such as creating new technologies, academic, journalism, research and development fields and so on. With today's technologies, it is possible to collect huge amount of opinionated data, such as speech, social media posts, political speeches, newspaper articles etc. and there is no limitation of processing these data in different ways. Processing all these text data is called Natural Language Processing. The process can be done with different methods, algorithms, programming languages, API's and packages.

In this project, there is an introduction to Natural Language Processing and there are some tools to analyze the text data. The purpose of this project is converting the voice data into text and taking selectable outputs, such as keywords, word frequencies, sentiment analysis. With these outputs, especially the sentiment output, throughout the whole opinionated data, a brief summary will be obtained about the consumer feedback, possible needs and the deficiencies.

# 1 Introduction

Consumer interaction is one of most important things in most businesses, hence companies and consumers needs to be in a communication before, during and after a sale and using operation to catch a certain qualification and take feedback about the defects of a product. It would be nearly impossible for employers to process and interpret all the data that came from the consumers. Therefore the NLP Algorithms get used by the firm for the feedback or possible developments. NLP is a very broad area to study on so it is very open for potential improvements such as search autocomplete, sentiment analysis, entity analysis, syntax analysis, chatbots virtual assistants etc.[1]
The purpose of this project was to process the data with certain NLP tools and algorithms that came from the customers, and interpret selectable different output under the consumer satisfaction topic.

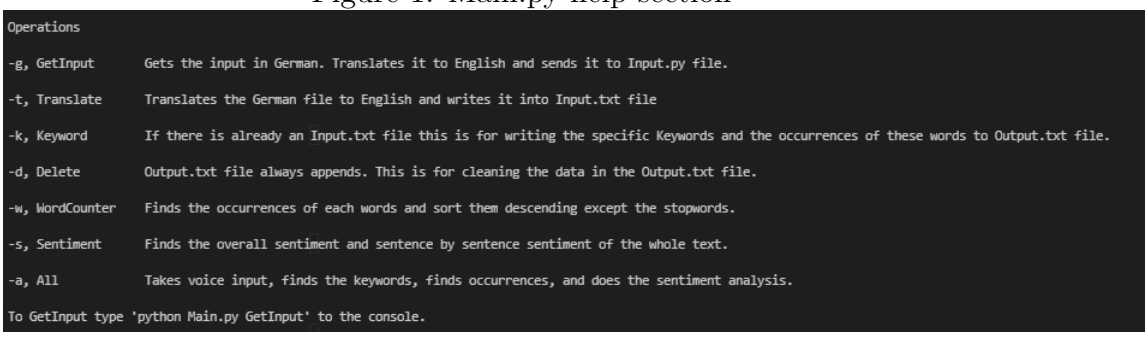
# 2 Methods

The project is coded as a command line project, so the user has option to choose between the processes through the console. 6 different files coded for the project.
GetInput.py coded in the beginning in order to obtain the data and write it into a file. This file was coded for a microphone input as speech in German and translates it into English and continue processing the data in English. SpeechRecognition package was used to obtain the speech input. SpeechRecognition package is available to use in several languages. For the translation Google Translate API was used.
Afterwards, Translate.py file coded in case of an already available German file. Translated the file in German to English and continued the process. Also GoogleTrans API was used in order to translate the input from German to English.
Third, KeyWord.py file coded to find the specific keywords throughout the text. It took the English input file and read the file as a start. Afterwards, the text tokenized into it's sentences and pushed into a list. Tokenization was made by nltk package. It could be written with algorithms but the packages work much faster that the hand written algorithms. After the sentences get pushed into lists, user asked for the amount of keyword that has been looking for and created a list for the keywords as size as the amount. Then, indexes of the lists are searched for the specific keyword that the user entered. The amount of the keyword stored in a variable. Keywords, frequencies of the keywords, sentences which the keywords occurred are written into an output file.
Subsequently WordCounter.py file coded in order to find the frequency of each word except stop words. First the text has tokenized to its words in order to separate the words from each other. Afterwards, stop words in English, such as "the", "a", "an", "that", etc., removed from the text in order to process it in a reasonable way. For this, nltk package used for it's stopwords library. Afterwards, the special letters removed from the text because word tokenizer has tokenized the special characters as an independent expression. Regex package used for its special characters library. Words are pushed into a dictionary and counted the frequencies with that dictionary throughout the text. Finally the frequencies of the words were written into a file by descending sort according to it's frequencies.

Afterwards, Sentiment.py coded. Sentiment analysis had high importance for this project because sentiment analysis is a very important analysis for product development, consumer recommendations and so on. The importance of the sentiment analysis will touched upon. Sentiment.py file is used for analyzing the input text sentence by sentence. It also used for writing the overall sentiment score as negative, positive, neutral. For sentence by sentence, the text also got tokenized by sentences and analyzed by it's polarity scores. NLTK Vader is used for the scores.
Finally, Main.py coded for a console project. The console project provides user the ease of choosing between different files. With "-h" command, the commands will show up to the page and the user will be able to choose the option.

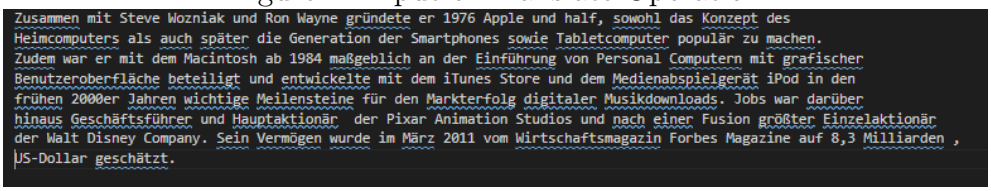
# 3 Results

Figure 1: Main.py help section



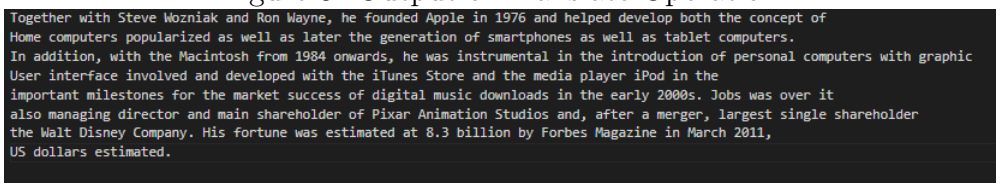

As it appears on the figure above, the project is user oriented and it provides the freedom of choice in case of already having some of the necessary files. To run GetInput file, it needs to written python Main.py GetInput to the console.

Figure 2: Input of Translate Operation




The original text for the translation is shown in the Figure 2 and the Figure 3 below shows the output of the translation. Figure 3 was the test data throughout the processes.

Figure 3: Output of Translate Operation

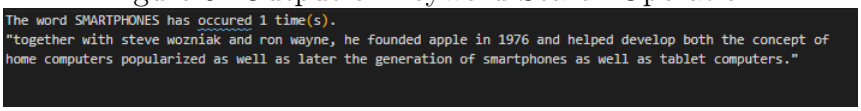



After the translation is done, it moves forward with keyword detection. First, algorithm asks about the amount of the keyword that the user demands, afterwards, it asks for the keywords. Afterwards, it found the

Figure 4: Input of Keyword Search Operation

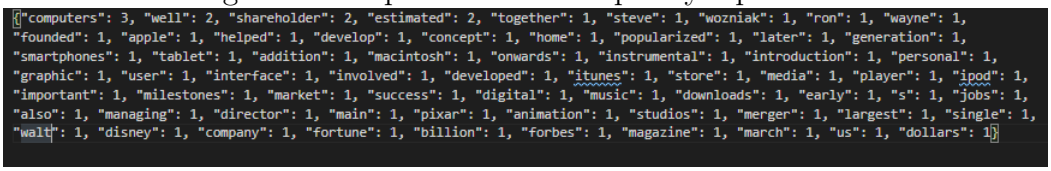

keywords and wrote them down in an output file which is:

Figure 5: Output of Keyword Search Operation



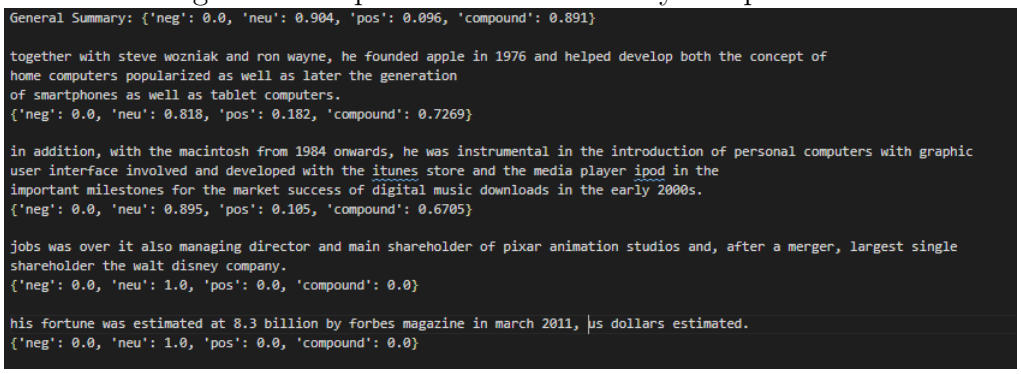
Afterwards, word frequencies are found.

Figure 6: Output of Word Frequency Operation




As it's shown on the Figure 6, all the words, except the stop words, are found by the algorithm and wrote in a descending form in order to finding the prior content.

Figure 7: Output of Sentiment Analysis Operation




As it is seen above, the algorithm calculated the sentiment of each sentence and wrote an overall summary to the whole text. As this is a text from wikipedia[2], this is an informative text so the sentiment is near neutral that is the reason of the high neutral values.

# 4 Sentiment Analysis

Sentiment analysis is a method to process a plain text data. The output of the analysis gives information about the emotions of the text. Today, there is a data overload for the customer feedback generally, so sentiment analysis is created to overcome the overload of customer feedback data. Sentiment analysis is used for gaining rational outcomes from very large text datasets. It provides to find the most important problem, and most concerned defect of a product/service or find the most satisfied part of a product/service and develop it.
Sentiment analysis can be made with different algorithms and packages. It is a semi-supervised algorithm generally. The algorithm usually learns from the social media, internet marketing comments. Compares the content of the comments with their rate of satisfaction. Sentiment categorizes the data under 3 different categories.

1. Positive
   Tests the positivity of the text/sentence and gives a value between 0 and 1.

2. Neutral
   Tests the neutrality of the text/sentence and gives a value between 0 and 1. This value is high in information texts.

3. Negative
   Tests the negativity of the text/sentence and gives a value between 0 and 1.

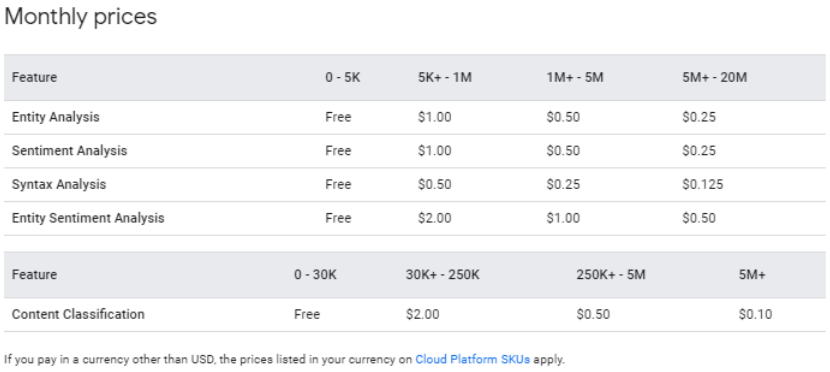
4. Overall
   Calculates the overall polarity score of the whole text according to the positive, neutral and negative values.

Score of these positive, negative, neutral and overall values are determined by the values that already exists for each word and there are also buffer words for multiplying the values with a certain constant value. As an example, "I like it" sentence takes a less positive value than "I like it very much" sentence. In this example "very much" is a buffer word which changes the overall value of the sentence[3].

# 5 Google Cloud Platform

Beneath the packages that are used in the project, Google has it's own cloud for Text Mining, Sentiment Analysis, Speech-to-Text, Text-to-Speech algorithms and these algorithms may have a higher accuracy. Also it has an advantage on the warranty and liability. It is generally used for data privacy, databases and data storage but also it has different API's for different usages such as Machine Learning, Deep Learning, Natural Language Processing, Data Mining. Cloud Natural Language is used for NLP and it has different prices for different usages.

Figure 8: Prices of the Natural Language API



Monthly prices

| Feature | 0 - 5K | 5K+ - 1M | 1M+ - 5M | 5M+ - 20M |
|---|---|---|---|---|
| Entity Analysis | Free | $1.00 | $0.50 | $0.25 |
| Sentiment Analysis | Free | $1.00 | $0.50 | $0.25 |
| Syntax Analysis | Free | $0.50 | $0.25 | $0.125 |
| Entity Sentiment Analysis | Free | $2.00 | $1.00 | $0.50 |

| Feature | 0 - 30K | 30K+ - 250K | 250K+ - 5M | 5M+ |
|---|---|---|---|---|
| Content Classification | Free | $2.00 | $0.50 | $0.10 |

If you pay in a currency other than USD, the prices listed in your currency on Cloud Platform SKUs apply.

Cloud Natural Language, empowers developers to easily apply Natural Language Understanding to the feature. There are dataset models in the cloud, such as big libraries for the words and the correlation between the words in a certain language, or the pronunciation of the word in a specific accent. This provides to process the data in German or easily translate it to English. As you can see above, there are different analysis options for the dataset. Entity analysis is used for labeling the content of the text, such as e-mail, chat, social media and with the sentiment analysis, shows the customer opinions to find actionable product and UX designs. Also there is a multimedia and multilingual support. Also the Cloud is able to classify the content such as sports, entertainment, healthcare etc. Also the Cloud offers us the same deep machine learning technology that Powers both Google Search's ability to answer specific user questions. This might be useful for a virtual assistant.

Also there is a cloud named as Dialogflow which provides the library for a chat bot and that is also can be used for a virtual assistant. And there are packages for processing the data in Google Cloud Platform[4].
The advantage of Cloud is, it has a high reliability and license agreements will not bother after the payment, there are lots of libraries in it and they can be used in any programming languages and it has one of the biggest modules for deep learning and natural language. The disadvantages of the Cloud is, it is paid and paid for every different item in it and also there might be some alternative packages in various programming languages.

# 6 Possible Developments

First of all, the project is coded as a command line project because first aim was to build a package and pushing it into pip, than it would be able to download it and free to use it. Also the Sentiment Analysis part can be developed as finding the most negative correlated sentence and sort the sentences descended by their negativity. Eventually, sentiment analysis values are gained by user comments and it is for developing the consumer satisfaction and finding the lacks of products in an autonomus way so the usage should be for research and development.
Besides, with a summarization algorithm, negative correlated sentences can be summarized and it would be easier for us to see the lacks of the softwares.
Finally, a virtual assistant or a voice command system might be created with these package libraries for the ease of use on the customer side.
Also the Google Cloud Platform might be used for a higher accuracy on translating and obtaining user input.

# References

[1] https://www.wonderflow.ai/blog/natural-language-processing-exampleswhy-use-nlp.

[2] https://de.wikipedia.org/wiki/Steve_Jobs

[3] https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664

[4] https://cloud.google.com/apis