PROJECT REPORT (Data Analyzer For WhatsApp Chats):

Text data analysis has emerged as a powerful tool for extracting meaningful insights from vast amounts of textual information. This report delves into the diverse possibilities of text data analysis and explores the interesting discoveries and applications that can be derived from this analytical approach, following by interesting technologies (libraries, applications, etc.) that have already taken a part on this environment from which we can learn and get inspired. Finally we will discuss briefly about the development process of our project, stating what we will try to implement and how we will tackle it.

Possibilities of analysis of text data:

Text analysis, also known as text mining, is the process of compiling, analyzing, and extracting valuable insights or information from large volumes of unstructured texts, using machine learning and NLP (natural language processing) techniques.

The sheer volume of data available on the internet today is incomprehensible. And manually analyzing this data is not really an efficient option. Just to be aware of the current situation, let's look at some numbers. In 2014, there were over 2.4 billion internet users either consuming or generating content. The number grew to 3.4 billion internet users by 2016. The coming year, 2017, added another 300 million internet users. 2017 also began the great information boom. Over 90% of information on the internet was created in or after 2017 and is mostly in textual format. The year 2020 recorded a total of 4.66 billion internet users.

There sure is a vast amount of information on the internet, and taking advantage of it for our own benefit is not a minor challenge.

All in all, taking into account this context lets see some interesting usages of this subject:

Topic Labeling: is a text mining technique that can help with categorization and interpretation of large volumes of textual data based on the theme of the information source.

Sentiment Analysis: also known as opinion mining, is a text mining model that can analyze and interpret the sentiments expressed by the author in any piece of text. In business ocntext, brands utilize sentiment analysis to assess public opinion by evaluating reviews and feedback from various sources. The analysis categorizes customer responses into positive, negative, and neutral sentiments, aiding brands in understanding their reputation. Examining negative feedback can pinpoint specific areas for targeted improvement.

Entity Extraction: text mining model consists on extracting different entities from any piece of text. The named entities are then classified as per pre-defined categories such as people, places, brands, monetary values, medical codes, and much more.

Text Classification: is based on categorizing text into predefined classes or labels streamlines in behalf of information organization. This is applicable in various contexts, from sorting documents and emails to filtering social media content, enhancing efficiency and user experience.

In general, there are as much usages of text data analysis as you can imagine, but we considered this examples are just enough to cover the basic ideas.

Tools and different technologies:

In order to accomplish this kind of tasks related with text analysis we count with several libraries and technologies that might be useful.

Natural Language Toolkit (NLTK):

Natural language toolkit is a suite of Python tools that allows users to analyze human language data using classification, tokenizing, tagging, and more. Though it requires some technical expertise to run, it can be very useful for both teaching and analysis.

Voyant:

Voyant is a web-based, dashboard-style tool that allows users to upload a corpora and visualize patterns in various ways. For instance, users can experiment with colorful word clusters that represent word frequency and visualize how specific words and phrases appear across texts in line graphs.

Mallet:

Mallet is a machine learning software program that is used through the command line with Python. Though it requires some technical skill to install and run, it can produce powerful results by generating “topics,” or lists of words that frequently appear together in corpora.

Gephi:

Gephi is a visualization tool that allows users to make colorful graphs and networks from textual data by revealing links between textual objects, social network patterns, and more. Gephi is easy to use and popular among humanists and social scientists alike.

The fact of being in this list does not mean that all of this technologies will be used in our Project, it is just general information of tools related with the topic. Either way, we will use lots of functionalities of these different programmes as inspiration for our own one, as there are very interesting features we can include.

Features to try out and development process:

For instance, we will Split our project development into two different stages which will be explained now, remarking different features we would like to implement and develop.

It has to be mentioned that we are going to be using 'Streamlit' which is a free and open-source framework to rapidly build and share machine learning and data science web apps. It is a Python-based library specifically designed for machine learning engineers, which is perfect for our goal.

1st Stage: for the beginning of the development some basic stuff will be tried and tested, nothing complicated such as simple statistics like; percentage of participation in a group chat, message density over time, etc. Just to get in tune with the framework and provide some basic but interesting data by means of plots (matplotlib for python). This will compound the first half of the project which will be delivered before the 20th December 2023.

2nd Stage: at this point we will already have a solid structure to continue working from, not because of that it is going to become easier. We will analyze and calculate more sophisticated statistics like topics in the conversations for example. To enhance our data analysis we will try to start using NLTK (Natural Language Toolkit) as it will make us reach the next level in terms of data mining, which is another way to refer to this.

In conclusion, data analysis is an enormous field and for sure it would be better to look at it more in depth, but in this project we tried to keep ourselves at the basics just to show the potential of it. Who knows, probably in the future we would take part in some much more sophisticated projects of this concrete subject, but already knowing the basic stuff thanks to jobs like this.