Ubottu - Abbas Pardawalla - Ubuntu Chatbot Creation Log

Non-technical overview of the subject area and the problem statement / opportunity identified

Using the publicly available dataset on Kaggle on the Ubuntu Chat logs, we hope to create a chatbot that aims to converse similar to a human.

Our bot will not only be able to chat and solve Ubuntu related problems, but also interact with an individual in a human like form. Thus creating opportunities for companies to create meaningful impact with its users and offering a much love 24/7 customer service

An overview of proposed vision for tackling the problem using Data Science

The powerful world of data science has made a massive leap with the tools in its arsenal. They have the ability to create chatbots, an Artificial Intelligence (AI) machine that is able to interact with people in a human like manner..

The tools that data science has for us to create these are, but not limited to :

- Python Interface to be able to write the code and tell the machine on how to carry the steps and information that will needed to be passed to the user

- Natural language processing (NLP) tasks, such as text classification, sentiment analysis, and language translation will assist the chatbot to understand what the user is asking, even when different sentences are used by different users for same question

- Machine learning models to improve their performance over time. By remembering past conversations and additional feedback from users, the model will be able to better itself

An estimate of the potential impact of such a solution

A Chatbot in new technical era has taken a massive leap - with introduction of OpenAI and CoPilot - population has seen the impact of its powers.

Over the years, the impact is surely set to only rise and help business and individuals in a multi factor of ways

Positive

1. Increased Efficiency - With that the increase of productivity
2. Enhanced Customer Experience - Reaching more customers on a global level
3. Cost-Effective - With less people and higher productivity
4. Accessibility - Not limited to the locale one is situated at, but reach everyone across seas

Negative

1. Job Losses - The work force needed will reduce or learn new skills
2. Dependence on Technology - New generation now knowing how to do basic skills. Loss of human abilities
3. Privacy Concerns - With all data able to everyone, ability to steal data will be simpler
4. Technical Issues - What happens when your internet goes offline

An introduction to the dataset, including data quality concerns and findings from preliminary EDA.

The Ubuntu Dataset is broken down into Three Unique Files

- dialogue_texts contain: 1038325 lines - 11035331 words & 116070597 characters
- diaglogue_texts_196 contain: 9212878 lines - 91660344 words & 996253904 characters
- diagoue_texts_301 contain: 16587831 lines - 166392849 words & 1799936480 characters

Each of the File also has Missing Values. Together combined

- Total Number of line missing texts are 2150. These will need to be deleted, as if there is no text, there is nothing for our model to learn.
- Total number of user info from our 'to' columns Is subsequently very high. Some cleaning is needed there or if able to match with known data

The Data in General seems clean and well recorded. The date, to & from column helps us navigate the most active users and the general vibe of the time the chats were logged.

The biggest concern for this dataset - is its size. One wrong code, and there goes the whole effort while running

Next steps in terms of data processing, feature engineering and baseline modeling.

The next step we will have to do after our preliminary EDA, is to take care of the problems we have encountered in our set.

- Data cleaning to make sure our columns are the right data type.
- Looking at missing values in the columns and dealing with them appropriately
- Choosing what to keep and remove from our dataset
- Finding what will create a positive/negative impact on our model

Through the process of adding new features in our code, we should be able to better our bot. By adding features from our first dataset, additionally the second and subsequently the third we can see how each feature will be able to better our code and bot. Initially the biggest challenge we have will is transforming categorical features into numeric ones and applying arithmetic operations to existing features.

The models we wish to create, there exists multiple other version in different classes. We can even use what exists there to better our model and learn from them.
In addition to all of this, deep learning will help us make our model stronger.