Trustpilot

Data Science Recrutiment Exercise Project Plan

Felipe Antonini Miehrig

1

1. Introduction

Trustpilot, as review platform, also receives thousands of reviews each year. Analysing this overwhelming amount of text manually proves itself to be a laborious task. Thus automatic ways of extracting insights are favored to support stakeholders in the decision making process. This small project aims to develop a simple tool to extract keywords and topics from user reviews and pair them with the appropriate business team.

2. Methodology

2.1. problem

In order to extract keywords and match sentences to topics (teams) it is important to represent words numerically. This can be done through word embeddings. Pre-trained Deep Neural Networks like Transfomers can facilitate this representation. By passing words and sentences as input to the models and extracting the weights of the last layer, we obtain a rich representation of the exert that can be compared to other words and sentences through cosine similarity.

2.2. Keyword extraction

There are already existing solutions to tackle the keyword extraction problem, like Keybert, that uses the encoder-based Bert class of LLMs to represent and then calculate the distance between tokens and the whole sentence, extracting keywords as a result.

2.3. Topic matching

After obtaining the N keyword in each review, we still have to pinpoint which team (topic) the review at hand is more associated with. For this, we use enriched representations of each team word. For instance, for the 'Legal' word, we wrap it into a string containing synonyms and related words for a more robust representation. We pass those and the reviews through the network extracting the embeddings and then calculating their cosine similarities. The team that yields the highest similarity score to the review is then assigned to it. For both tasks, we use a multilingual version of Bert model to allow comparisons between different languages.

3. Presentation

The results are saved into a data frame with the same layout as the original data. N new columns are added, each containing a "keyword" for the review. Additionally, one column is added for each team showing the degree of pertinence (cosine similarity) of the team definition and the review content. The team with the highest score is displayed in a final column as "related team". Plots are also provided. Alongside charts analyzing the original dataset, plots containing the newly mapped number of reviews for each team per week are shown. A simple quality control methodology is used to indicate trends and whether points are outside the norm. A plot for keywords for each team is also provided, this shows the star rating on the horizontal axis and the pertinence to the team on the vertical axis.