Model Outcome

Team 1 SDGs

INTRODUCTION

The model built analysed each subcategory of SDG #3 singularity, to then put together the results and draw the final conclusion. The model accuracy was 63.33%, but the Team recognizes that a bigram tokenization technique could bring the accuracy higher in the future. The Team has also built a Prototype using html to provide an initial idea of how it will look like in the future.

METHODOLOGY

Before putting hands on the dataset and starting coding, the team approached the subject widely by reading and studying it more in depth. This helped in developing determined knowledge around Multi-Label Text Classification. The main source was the book 'Natural Language Processing with Python' by Steven Bird, Ewan Klein & Edward Loper.

It was right for the team to read and learn about the new skills needed so that the project could run smoothly. Moreover, the team collaborated in the data cleaning and preparation and exploratory data analysis.

In terms of data wrangling, we lowercased the data and removed stopwords. We tokenized the data for accessibility of the various functions we used.

We split the text into tokens and assigned a token to each of their respective labels. We computed the term frequency matrix and fed them to a Naive Bayes Classifier from the NLTK library for each SDG label. The Naive Bayes classifier compares the token with previously trained popular token and compares the new text. We then tested this algorithm on a test data of 100 random different pieces of text that none of the classifiers had seen before.

BUILDING THE MODEL

We tackled the classification problem using the 'bag of words' approach. Splitting the text into categories per labels and generating a term frequency table. Harnessing the probability of more frequent words by category would split the various SDG goals.

Each SDG label had their own distinct popular words for example, the goal 3.b.2 was related to medical research medical terms such as 'neuronal' were prevalent.

We broke each of the 27 labels into their own categories and made a binary classifier for each. We then joined the outputs of all these classifiers making it a multiclass, multilabel algorithm.

RESULTS

The accuracy of the model is 63.33%. Considering we do not penalise our model for getting extra labels. The reason for this choice is the time constraint not allowing us to tweek the sensitivity of the model.

PROTOTYPE

The following is a link to our prototype website coded using HTML:

https://veritext.glitch.me

KEY LEARNINGS

NLTK is a leading library operational since 2006. The library has all comprehensive tools required in textbook natural language processing. It is an out-of-the-box solution for most Text Analytics applications.

Naive Bayes classifier, albeit not as accurate as other available classifiers provides a good trade-off between speed and accuracy. It is a very powerful tool to use for preliminary modelling.

We concluded that bigrams, where two words instead of one are displayed, could be done. This would provide more context and more explanation for the words to be grouped more appropriately which can improve the accuracy of the model in the future.

CROWDSOURCING

Crowdsourcing is a sourcing model in which individuals or organizations obtain goods or services, including ideas, voting, micro-tasks and finances, from a large, relatively open and often rapidly evolving group of participants (Wikipedia). From the name “Crowd” and “sourcing”, the process is usually open with little to no limit to the pool of respondents to the process.

If well harnessed, crowdsourcing can provide the following benefits to business:

- A diverse pool of resources (expertise, ideas, finances). This will provide solutions to tough business challenges.

- A diverse pool of thinking
- Low cost of financing for businesses
- A rich source of data for business analysis
- Cheap source of marketing (brand ambassadors)

Based on the benefits above, we believe crowdsourcing should be exploited by the business. Although, it is not all benefits as there are downsides to crowdsourcing like danger of manipulation and danger of loss of image. These risks should be mitigated.

OFFICIAL SUBMISSION OF THE CHALLENGE

Jupyter Notebook containing the model will be submitted with this document as a proof. The competition was already closed so we could not submit it on the official website.

However, please find the link for the official challenge below:

https://zindi.africa/competitions/sustainable-development-goals-sdgs-text-classification-challenge/leaderboard

REFERENCES

Alexis Fournier, '6 Great Advantages of Crowdsourcing you can benefit from'. Retrieved from https://www.braineet.com/blog/crowdsourcing-benefits/

Crowdsourcing, retrieved from https://en.wikipedia.org/wiki/Crowdsourcing

Steven Bird, Ewan Klein & Edward Loper, 'Natural Language Processing with Python'