

Project Two

Dsc 680



October 15, 2022

Colton Cook

**Table of Contents**

**Business Problem………………………………………………………………………………………………………………….2**

**Background/History………………………………………………………………………………………………………………2**

**Data Dictionary……………………………………………………………………………………………………………………..2**

**Data Explanation……………………………………………………………………………………………………………………2**

**Methods………………………………………………………………………………………………………………………………..3**

**Analysis…………………………………………………………………………………………………………………………………3**

**Modeling Setup……………………………………………………………………………………………………………………..6**

**Conclusion…………………………………………………………………………………………………………………………….6**

**Assumptions………………………………………………………………………………………………………………………….6**

**Limitations…………………………………………………………………………………………………………………………….7**

**Implementation Plan……………………………………………………………………………………………………………..7**

**Challenges……………………………………………………………………………………………………………………………..7**

**Future Uses/Additional Applications……………………………………………………………………………………..7**

**Ethical Assessment………………………………………………………………………………………………………………..7**

**Business Problem**

Applying topic modeling to documents or any collection of text is a great way to categorize the data. From a business perspective in ads, it would allow data to be organized by the various topics the data contains and would allow ads to be targeted at the data for the reader to then see. It a model can categorize the data accurately and into generalized topics, it would be a great way to target ads.

**Background/History**

Solving the accurately of topic modeling and being able to do it fast enough to allow real time labeling would be a great step forward for companies that depend on it, like most social media companies and search engines. Having done some projects on this topic in the past, it was hard to get high accuracy of the labeling even with a human it was difficult. Some advances in the field have improved the process for models and labels incredibly high. Having topics that match up well with

**Data Dictionary**

|  |  |
| --- | --- |
| **Feature** | **Description** |
| publish\_date | Date article was published |
| headline\_text | Headline text of article |

**Data Explanation**

This data was made available from an Australian news company I believe a few years ago.

Text

Description automatically generated with medium confidence

**Chart, histogram

Description automatically generated**

**Methods**

Exploratory data analysis was done using python. Used typical python libraries for this process in pandas, numpy, matplotlib, seaborn. Fairly typical process for exploring the data. Modeling was handled with Bertopic. A custom BERT model that is optimized for topic modeling someone created and open sourced it. All the code was done in a jupyter notebook.

**Analysis**

After cleaning up the data and getting it into a tubular format, I decided to look at various distributions of the data around some of the features the model found.

The below table has 3 main columns, providing information about all the topics in descending order of topics size/Count.

* 'Topic' is the topic number, a kind of identifier, and the outliers are labeled as -1. Those are topics that should be ignored because they do not bring any added value.
* 'Count' is the number of words in the topic.
* Name is the name given to the topic.

Table

Description automatically generated

For each topic, we can retrieve the top words and their corresponding c-TF-IDF score. The higher the score, the most relevant the word is in representing the topic.

Graphical user interface, text, application, email

Description automatically generated

From this topic, we observe that all the words are coherent for the underlying topic which seems to be about firefighters.

The most relevant words of each topic can be visualized in a form of bar chart out of the c-TF-IDF score, which is interesting to visually compare topics. Below is the corresponding visualization for the topic 6 topics.

Chart

Description automatically generated

For those who are familiar with Latent Dirichlet Allocation LDAvis library. This library provides the user with an interactive dashboard showing for each topic the corresponding words and their score. BERTopic does the same with its visualize\_topics() function and even go one step further by giving the distance between topics (the lower the most similar), and all of this with a single function visualize\_topics()Chart, scatter chart

Description automatically generated

As you can see in the Intertopic Distance Map dashboard, some topics are very close. One thing that could come to mind is how can I reduce the number of topics? The good news is that those topics can be hierarchically in order to select the appropriate number of topics. The visualization flavor helps to understand how they relate to one another. This next chart contains 30 topics.

Table

Description automatically generated with medium confidence

**Modeling Setup**

The modeling itself is easy to setup as its all contained in the bertopics library. The model will run on most data with the right setup. Its important to look at the graphs and charts to make sure the model is acting accordingly and giving appropriate results.

**Conclusion**

The model was able to label and categorize the data very well. Having checked a few of the topics, it seemed very accurate and made sense in most cases. There were a few that were overly board or generalized. Under some occupations it listed human. Seemed a little weird, but it wasn’t wrong.

In the end, it seemed good enough to be used for labeling documents and could be used in the ad space or other industries that need to label data in text format.

**Assumptions**

With all data, I am assuming it is accurate and correct. With this model and method of labeling, it does have some issues with reproducibility. That can cause some issues if you aren’t careful. Making sure a random seed is chosen and applied can be important.

**Limitations**

I didn’t check every label. So, the topics could be off in some of them. It’s also possible that labels would need to be changed based on the content if it had a low of varying information. I could imagine a sci fi blog could have a lot of weird topics generated and someone would have a harder time trying to put ads or labels that make sense without customizing the model more.

**Implementation Plan**

This model could be deployed in production. It runs fast and is efficient. It also doesn’t need to be run often. New data could be added to it quickly and it could be labeled. Deploying this in the cloud or even locally would be easy to do and not require a lot of technical buildup for it.

**Challenges**

There are no real challenges with this model setup or running it.

**Future Uses/Additional Applications**

Could be refined more to include more precise topics, but for an out of the box model it runs very well.

**Ethical Assessment**

There are some issues. First one being is the text monitored first. If not, bad things could slip through and be given labels that aren’t correct or influence other labels. Second, is the data being labeled correctly? Was an article written in a way that gets labels that you wouldn’t want there and the customer paying for the ad targeting is mad because their ads are showing up for something they don’t endorse.