By Shreyas Nanaware

Project analysis

**Approaches Used**

1. Machine Learning Algorithms

Initially, the data was trained using all the machine learning classification algorithms. Tf-Idf vectorisation and Doc2Vec feature space was used to feed in into the model as training data.

Models used:

* Logistic Regression
* MultiNomialNB (Naïve Bayes)
* LinearSVC (Support Vector Classification)
* RandomForestClassifier

1. Neural Network

After evaluating the data on machine learning algorithms, the text data was trained on Neural Network like ANN (Artificial Neural Network) and

1-D CNN (Convolution Neural Network).

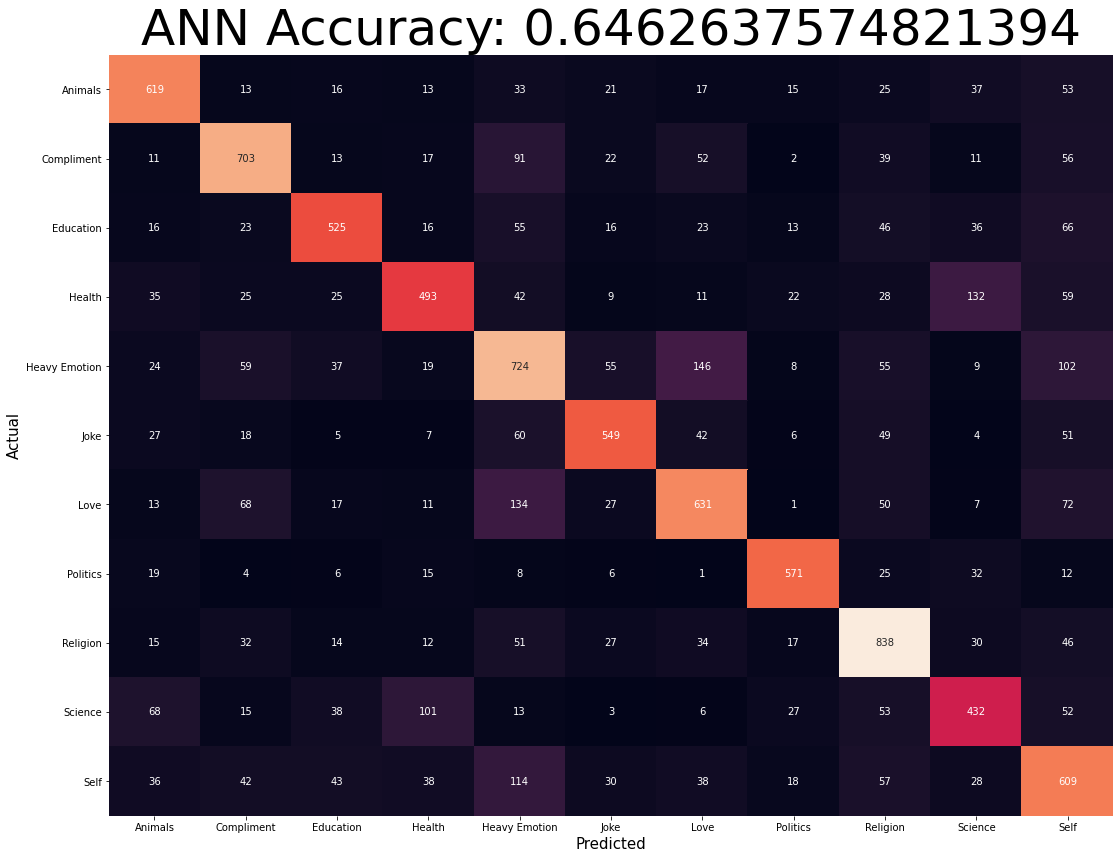
The data was also trained on BERT (Bi-directional Encoder Representation from Transformers) . It is an open source machine learning framework for natural language processing (NLP).

Model Comparison

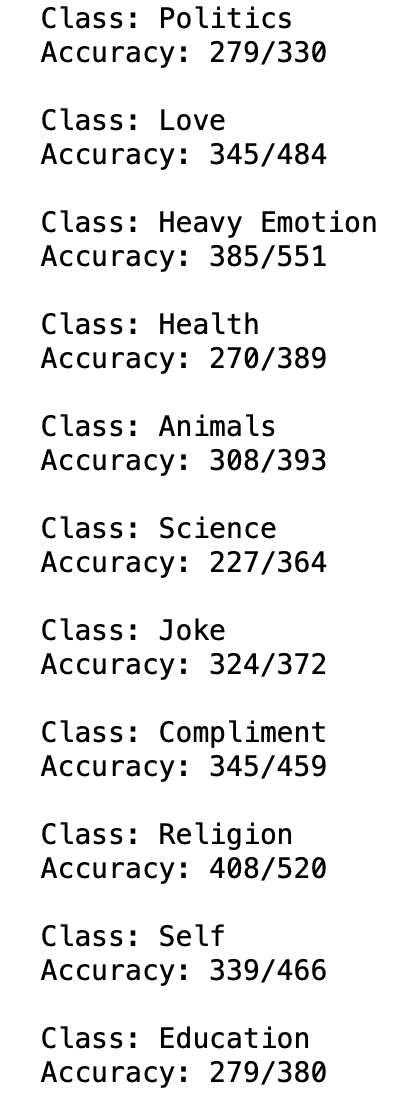
Machine Learning Algorithms

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Logistic Regression Classification Metric |  |  |

Neural Network

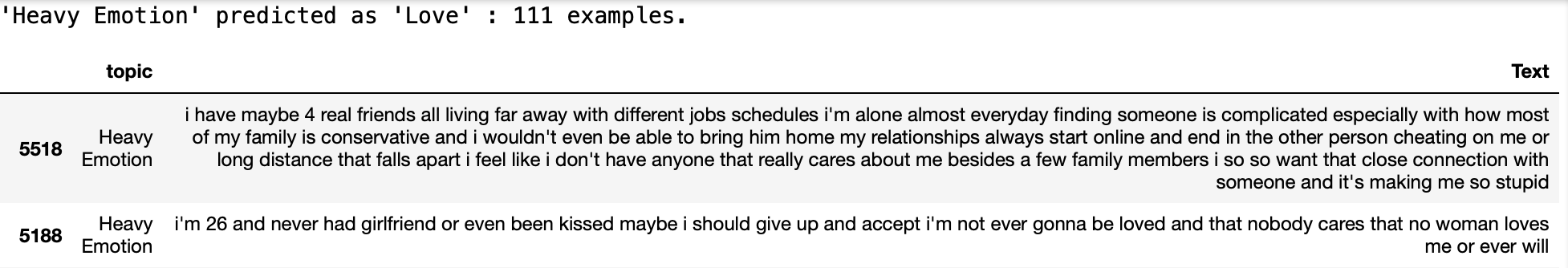


**BERT** – Model was giving an accuracy of 74.5 % on the validation set.



Analysis

* Topics like ‘*Heavy Emotion’* and ‘*Love’* are very closely related.



The above example is where topic “Heavy Emotions” were predicted as “Love”. As the sentences in both topics are quite similar, the model is unable to distinguish between the two topics.

Possible Solution: Adding more data.

* Td-Idf vectorizer was used ahead of Word2Vec, as it was giving better accuracy. But generally Word2Vec embedding are more useful in understanding the semantic relationship between words and can work with small length sentences also, whereas in Tf-Idf, working with small sentence tends to create bias sometimes.

This issue is seen when topic ‘*Animals’* were classified as ‘*Love’.*

Possible solutions: Include sentences that has length greater than X (value after some analysis).

* Hyper-Parameter tuning is another approach that can be used in BERT to improve the model performance.
* As the value counts for all the topics has less variance, there was no need to perform Up-sampling or Down-sampling. In general more data is required for better training of neural network models like BERT.

