Introduction

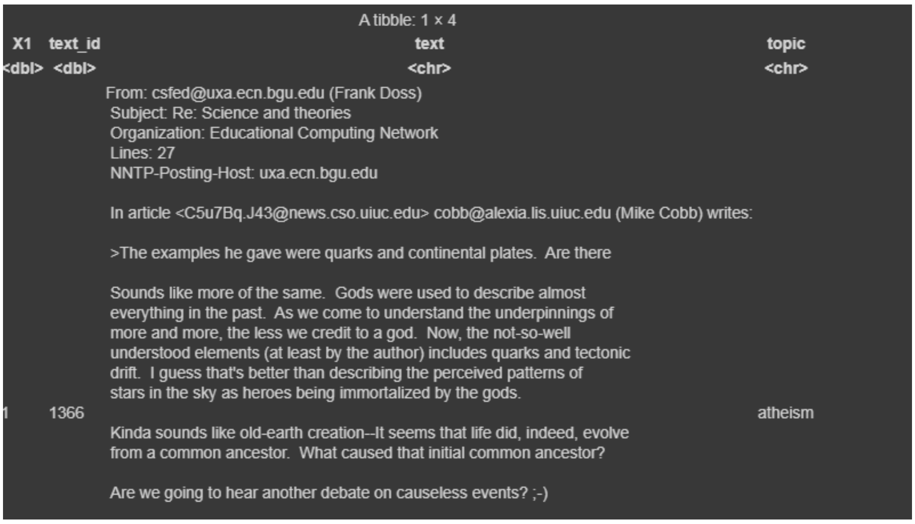
This project will employ machine learning to classify different newsgroups data and explore additional trends and observations. The dataset used is the well-known 20 Newsgroups dataset, which has become quite frequency used in the field of machine learning.The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It was originally collected by Ken Lang, probably for his Newsweeder: Learning to filter netnews paper, though he does not explicitly mention this collection. The reasoning behind choosing newsgroups data is mainly connected to also having real life applications. Through an advanced neural network it is possible to classify a certain piece of news simply by its content. This can be applied when uncertainty occurs about what category certain news might belong to. In this case the journalist can use the algorithm to help determine what specific category it might belong to. Additionally the algorithm can be used to suggest another category which can make the news fall into multiple categories. This can save time from various journalists/news publishers, who are no longer required to manually determine what category certain news belong to. This project will start by exploring the data with the use of a Glove embedding model, where the link between different words is being researched. Afterwards a network analysis will be constructed to explore how pair of words are connect with the use of Bigrams. After exploring the text a baseline Random Forest model will be constructed to see how well it performs.
Subsequently a neural network will be constructed with the goal of classlifying multiple topics. The neural network used in this project is "Long Short Term Memory", which is a superior neural network for sequential text data. Lastly the project will conclude upon the discoveries made.

The data used in this project is the well-known 20 Newsgroups dataset. It contains 18773 observations containing newsgroups post, which are split into different topics/categories. The 20 different topics are:

- Atheism
- Autos
- Baseball
- Christianity
- Comgraphics

- Comp-windows
- Crypt
- Electronics
- Guns
- Hockey
- IBM-PC-hardware
- Mac-hardware
- Medicin
- Mideast
- Misc-forsale
- Motorcycles
- Politics
- Religion-misc
- Space
- Windows-x

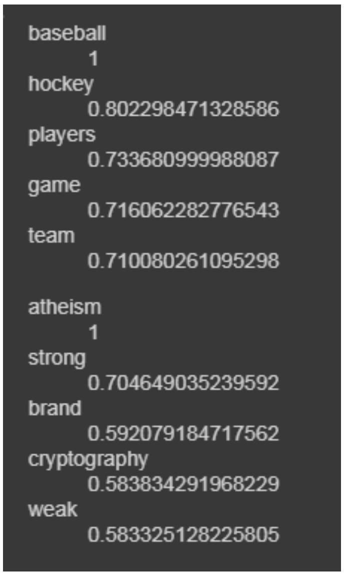
An observation from the dataset is presented below and the data is also publicly available on http://qwone.com/~jason/20Newsgroups/.



```
                                  A tibble: 1 × 4
 X1   text_id                          text                                    topic
<dbl> <dbl>                            <chr>                                   <chr>
             From: csfed@uxa.ecn.bgu.edu (Frank Doss)
              Subject: Re: Science and theories
              Organization: Educational Computing Network
              Lines: 27
              NNTP-Posting-Host: uxa.ecn.bgu.edu

              In article <C5u7Bq.J43@news.cso.uiuc.edu> cobb@alexia.lis.uiuc.edu (Mike Cobb) writes:

              >The examples he gave were quarks and continental plates.  Are there

              Sounds like more of the same.  Gods were used to describe almost
              everything in the past.  As we come to understand the underpinnings of
              more and more, the less we credit to a god.  Now, the not-so-well
              understood elements (at least by the author) includes quarks and tectonic
              drift.  I guess that's better than describing the perceived patterns of
              stars in the sky as heroes being immortalized by the gods.
1      1366                                                                    atheism
              Kinda sounds like old-earth creation--It seems that life did, indeed, evolve
              from a common ancestor.  What caused that initial common ancestor?

              Are we going to hear another debate on causeless events? ;-)
```

In the next section we will investigate some connections in the dataset, before actually making models with the option of making predictions. To accomplish this we will be using a word embedding model called a GloVe model.

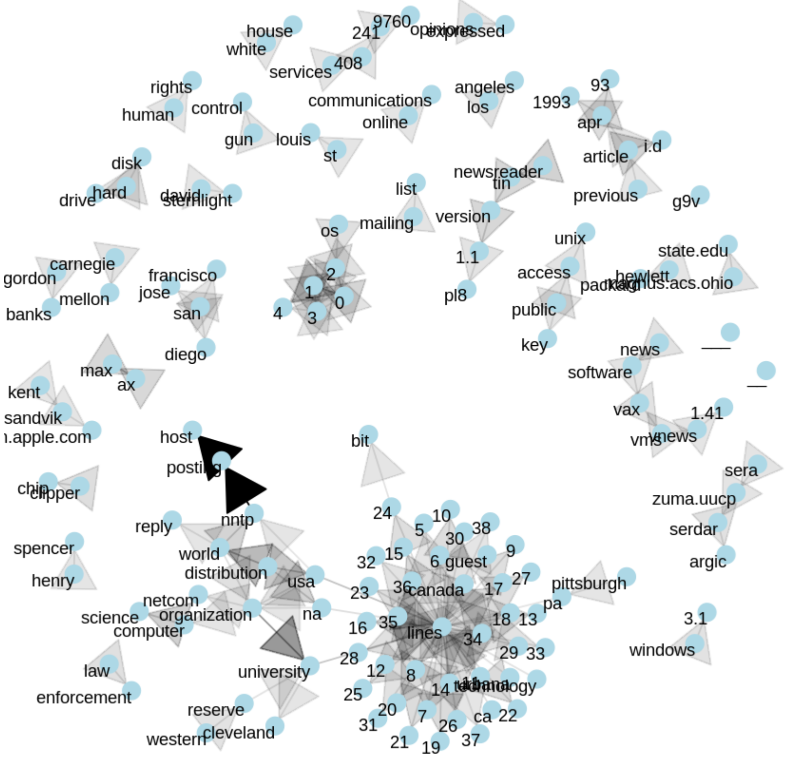
GloVe embedding model



```
baseball
         1
hockey
         0.802298471328586
players
         0.733680999988087
game
         0.716062282776543
team
         0.710080261095298

atheism
         1
strong
         0.704649035239592
brand
         0.592079184717562
cryptography
         0.583834291968229
weak
         0.583325128225805
```

The 5 most relevant words with the greatest correlation can be seen here. For baseball, it is seen that the word 'hokey' has the greatest correlation and then 'players'. For the subject of Atheism, the word 'strong' is the word with the highest relation. It is coded so that the top 5 words with the greatest correlation in a decreasing approach are displayed. This makes some sense in relation to the topic and what words are correlated with it.

Network Analysis

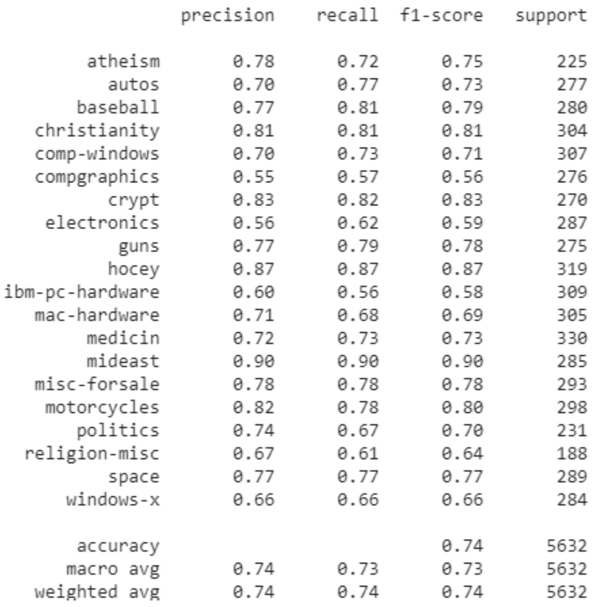


The plot above shows how the words connect in the newsgroup. Arrows from one word to another indicate that there is a connection. Since the words are lemmatized, it will not be grammatically correct, but it still makes sense to look at the connection. There are big and small connections, for example, law -> enforcement, computer -> science and white -> house small connections, where san -> Diego -> Francisco -> Jose is a larger connection with several words combined to each other. At the same time, a much larger cluster is seen around the word 'lines', which is the key word for all the numbers pointing towards each other. This makes sense since numbers are similar and therefore, they are interconnected.

Benchmark model

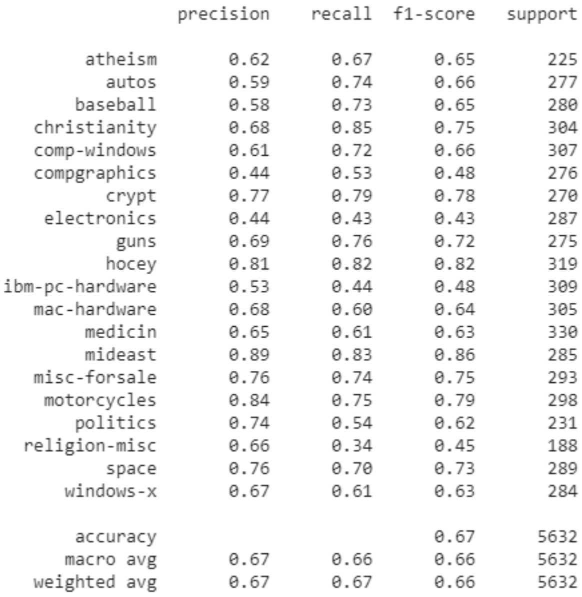
We are using random forest and a logistic regression model to hold our eventual neural network up against. The random forest utilizes decision trees to make calcifications and chooses the best path to make the final decision. The logistic model uses linear models to make predictions using a nearest neighbors approach to make decision, getting the following results.

Logistic Regression



```
                 precision    recall  f1-score   support

         atheism       0.78      0.72      0.75       225
           autos       0.70      0.77      0.73       277
        baseball       0.77      0.81      0.79       280
    christianity       0.81      0.81      0.81       304
    comp-windows       0.70      0.73      0.71       307
    compgraphics       0.55      0.57      0.56       276
           crypt       0.83      0.82      0.83       270
     electronics       0.56      0.62      0.59       287
            guns       0.77      0.79      0.78       275
           hocey       0.87      0.87      0.87       319
ibm-pc-hardware       0.60      0.56      0.58       309
    mac-hardware       0.71      0.68      0.69       305
         medicin       0.72      0.73      0.73       330
         mideast       0.90      0.90      0.90       285
    misc-forsale       0.78      0.78      0.78       293
     motorcycles       0.82      0.78      0.80       298
        politics       0.74      0.67      0.70       231
   religion-misc       0.67      0.61      0.64       188
           space       0.77      0.77      0.77       289
       windows-x       0.66      0.66      0.66       284

        accuracy                           0.74      5632
       macro avg       0.74      0.73      0.73      5632
    weighted avg       0.74      0.74      0.74      5632
```

Random forest



```
                 precision    recall  f1-score   support

         atheism       0.62      0.67      0.65       225
           autos       0.59      0.74      0.66       277
        baseball       0.58      0.73      0.65       280
    christianity       0.68      0.85      0.75       304
    comp-windows       0.61      0.72      0.66       307
    compgraphics       0.44      0.53      0.48       276
           crypt       0.77      0.79      0.78       270
     electronics       0.44      0.43      0.43       287
            guns       0.69      0.76      0.72       275
           hocey       0.81      0.82      0.82       319
ibm-pc-hardware       0.53      0.44      0.48       309
    mac-hardware       0.68      0.60      0.64       305
         medicin       0.65      0.61      0.63       330
         mideast       0.89      0.83      0.86       285
    misc-forsale       0.76      0.74      0.75       293
     motorcycles       0.84      0.75      0.79       298
        politics       0.74      0.54      0.62       231
   religion-misc       0.66      0.34      0.45       188
           space       0.76      0.70      0.73       289
       windows-x       0.67      0.61      0.63       284

        accuracy                           0.67      5632
       macro avg       0.67      0.66      0.66      5632
    weighted avg       0.67      0.67      0.66      5632
```

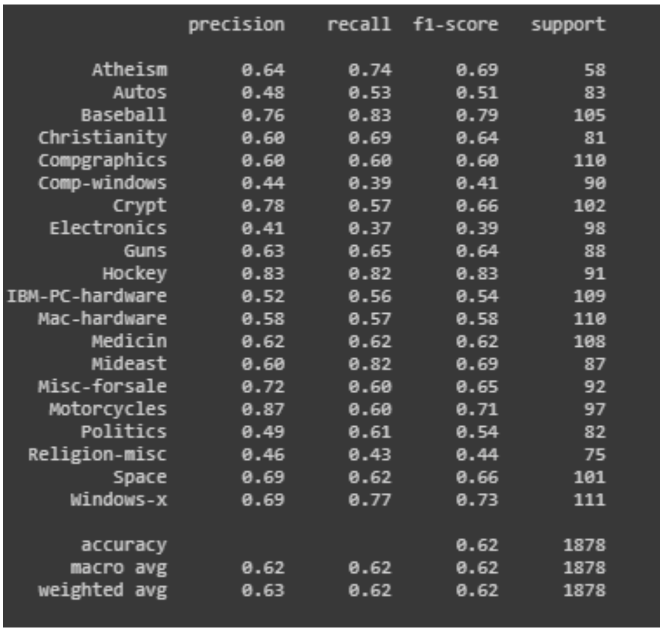
The most important stats above is the accuracy stat, with logistic regression having a 74% accuracy and Random forest having a 67% accuracy.

LSTM neural network

Long short_term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is commenly used for activities such as unsegmented, connected handwriting recognition, speech recognition and anomaly detection in network traffic or IDS's (intrusion detection systems).

LSTM models have been used in many different areas of machine learning. For example in 2018 OpenAI developed bots able to humans in the game of Dota 2. OpenAI Five consists of fiveindependent but coordinated neural networks. Each network is trained by a policy gradient method without supervising teacher and contains a single-layer, 1024-unit Long-Short-Term-Memory that sees the current game state and emits actions through several possible action heads.

Also in 2019 Deepmind made a program called "AlphaStar" that used a deep LSTM core to play the complex game "Starcraft. The results are presented on the following page.



```
                precision    recall  f1-score   support

       Atheism       0.64      0.74      0.69        58
         Autos       0.48      0.53      0.51        83
      Baseball       0.76      0.83      0.79       105
  Christianity       0.60      0.69      0.64        81
  Compgraphics       0.60      0.60      0.60       110
  Comp-windows       0.44      0.39      0.41        90
         Crypt       0.78      0.57      0.66       102
   Electronics       0.41      0.37      0.39        98
          Guns       0.63      0.65      0.64        88
        Hockey       0.83      0.82      0.83        91
IBM-PC-hardware       0.52      0.56      0.54       109
  Mac-hardware       0.58      0.57      0.58       110
       Medicin       0.62      0.62      0.62       108
       Mideast       0.60      0.82      0.69        87
  Misc-forsale       0.72      0.60      0.65        92
   Motorcycles       0.87      0.60      0.71        97
      Politics       0.49      0.61      0.54        82
 Religion-misc       0.46      0.43      0.44        75
         Space       0.69      0.62      0.66       101
     Windows-x       0.69      0.77      0.73       111

      accuracy                           0.62      1878
     macro avg       0.62      0.62      0.62      1878
  weighted avg       0.63      0.62      0.62      1878
```

We can observe from the above figure that the model can easier classify certain topics compared to others. For example the topic "Hockey" has a accuracy of 83%. The LSTM model experiences problems when having to classify Electronics, which can be explained through the other topics. There is so many topics related to electronics for example Windows X, MAC hardware, IBM-PC-hardware and compgraphics, which makes this topic and other topics hard to classify. IBM-PC-hardware and Mac-hardware accuracy is also quite low on 52% and 58%. Same for Comp windows on 44%, which indicates the model mixes the different topic with each other. Baseball is good example of a topic that is not related to many other topics. This is also where the model achieves very high accuracy. Same for Motorcycles and so on. Overall we can conclude the LSTM neural network is quite accurate, but more accurate for certain topics.

Conclusion

For this conclusion we will be focusing primarily on our benchmark models and the neural network, the Glove model helps us with identifying words that are important in each topic. We find that electronics are connected as well as sport topics, and that some combinations of words are used in specific categories.

Looking at the 3 models, we see that the logistic and random forest models are outperforming our neural network. While not expected, this makes some sense as the classification models are based on a bag of word approach making the words everything and as there will be words used exclusively in each topic. The neural network takes sequences into a count, the sequence of words is often a personal thing and will therefore work well in identifying a person from writing but less in identifying a topic. Further though our model has been run for some time it still has room for improvements, both by adding epochs and more layers and units in each layer and therefore there are room for improvements.