**Question Classification: What people are asking about Covid?**

**GitHub Repository Link:**

**Problem Statement:**

Classify the questions asked related to Covid into different classes or categories using COVID-Q dataset - <https://github.com/JerryWei03/COVID-Q>

**Solution to problem:** Text data needs to be converted into numerical data. Hence, for this embedding/feature generation we explore different methods like Tf-idf and word2vec in NLP.

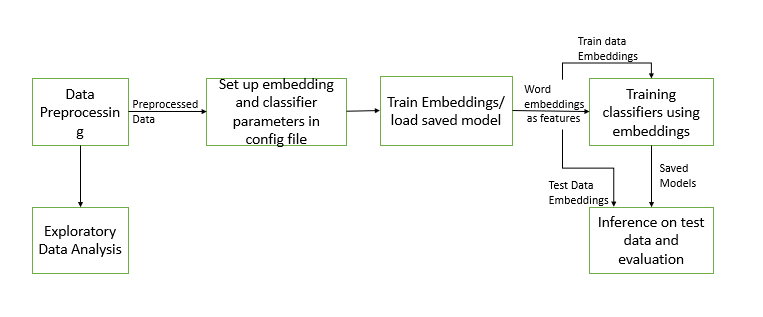
We will be comparing the performance of different classifiers like Support Vector Machine (SVM) and MLP (Multi-layer perceptron) with both the embeddings. We specially selected these two techniques as they work with high dimensional data like ours

**Details of the Dataset:** number of classes – 16, we have questions generated from different sources which we will be diving into train and test dataset. There are questions that are generated by authors. We keep all those for testing. The percentage of train test data division can be adjusted in the code. Also, whether to augment the data or not is set up in code.

Train, test data statistics (number of questions) using 30% split - train data: 293, 2, test data: 691, author test data: 249

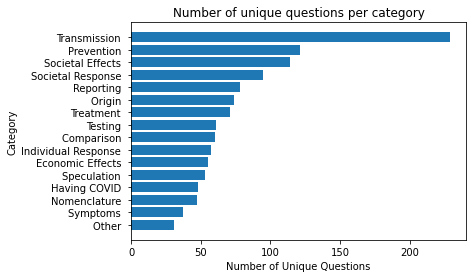
*Note: Please check – “ReadMe – How to run the scripts” file for more details.*

**Project Pipeline:**

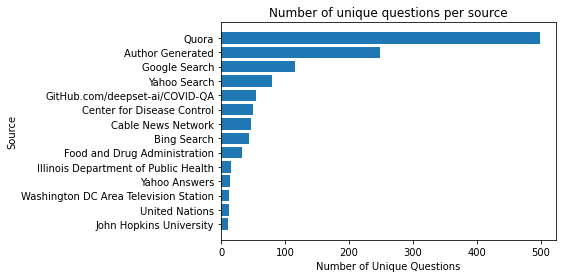
****

**Important Insights on Data: Using Exploratory Data Analysis (EDA)**

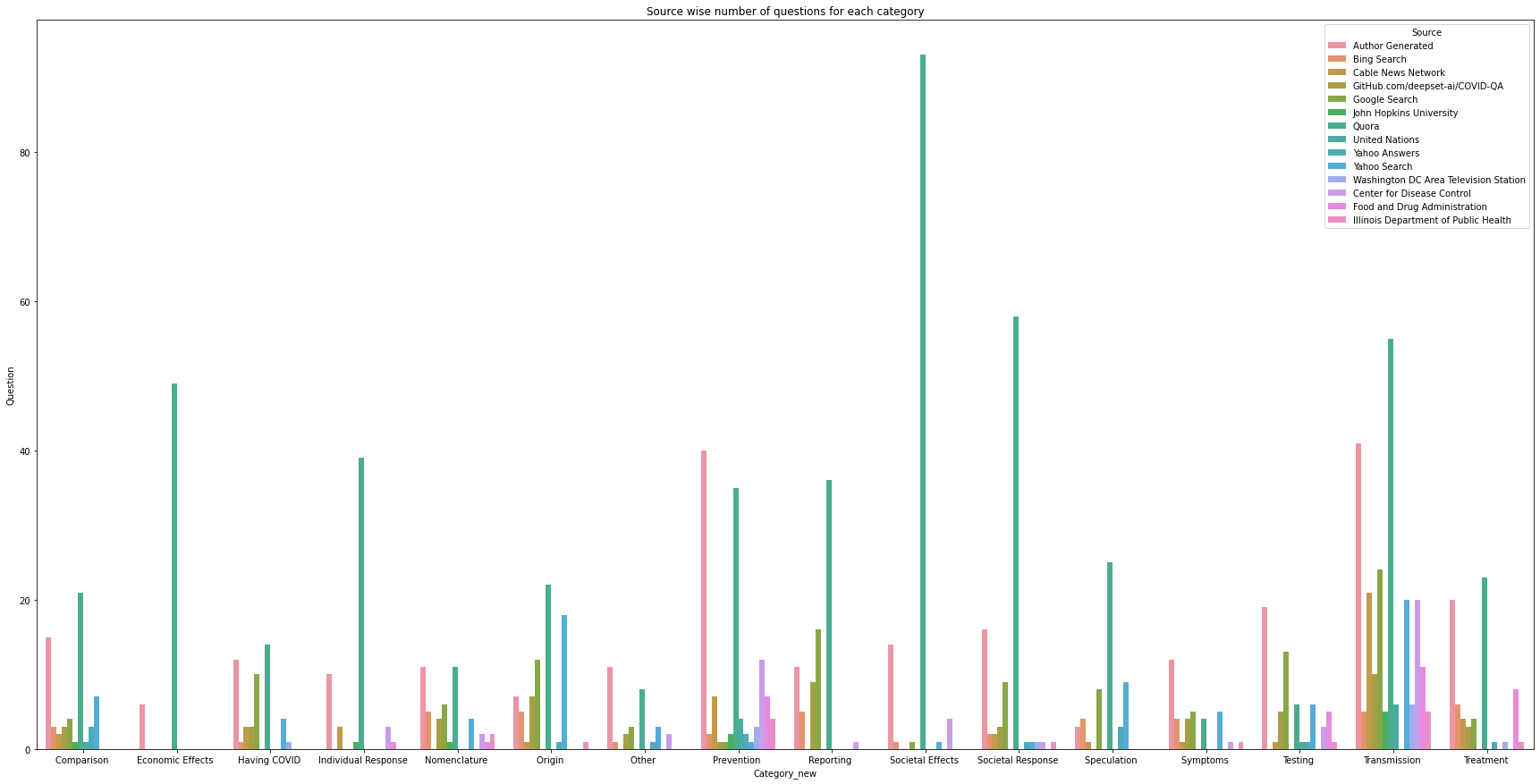
Number of questions for each category: Transmission has highest number of questions followed by Prevention



Number of questions for each source: Quora and Author generated most number of questions

****

Number of questions per category and sources



A picture containing text, newspaper

Description automatically generatedWordcloud for main categories: More frequent data with larger font.

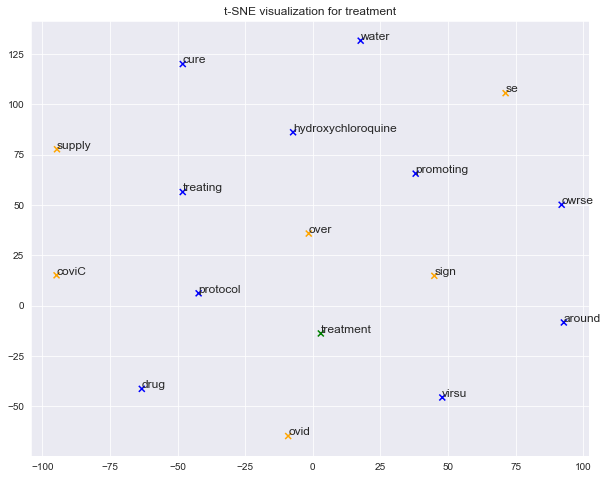
As we can say covid, can, spread, etc are main words as expected in transmission category

Text

Description automatically generated

Prevention has main words like vaccine, should, sanitizer, hand, etc.

Word2vec T-SNE visualization: Words that are similar are closer to main word for which t-SNA plot is plotted (ex – here for word treatment)



**Data augmentations performed on Train data:**

1. Char Level Augmentations
   1. KeyboardAug: It will replace random characters by keyboard distance, i.e. character will be replaced by item closer to itself in keyboard
   2. RandomCharAug: Applies random character error to textual input by replacing, deleting, inserting, and swamping random characters
2. Word Level Augmentation
   1. SynonymAug: According to wordnet, random word is replaced by the word with similar meaning
   2. RandomWordAug: Random word is deleted in the sentence
   3. SpellingAug: Inserts spelling error in any random words
   4. SplitAug: Splits any random words

Not all categories are augmented same. Number of questions are large for Transmission category hence only word level augmentations are applied for it to maintain class balance (similar number of samples for each category)

Number of questions in train data – 237, Number of questions in augmented train data – 288**7**

We use different parameters for each of classifiers (SVM- Support Vector Machine and MLP- Multi-Layer Perceptron) and with both word2vec and tf-idf feature generation techniques to come up with the best model that gives best accuracy.

Accuracy = (number of questions correctly classified)/(number of questions correctly classified + + number of questions incorrectly classified)/100

**Results:**

Tf-idf features generated – 3880, word-embedding dimension used - 700

On test data – SVM(linear classifier with balanced class) and tf-idf gave best results. This was expected as word2vec needs large dataset and questions sentences are short

**Table

Description automatically generated**

On author generated test data- MLP(neurons with layers – 64,32) and tf-idf gave best results

Table

Description automatically generated

**Challenges:** Data is very imbalanced, some of the question types have very less data and hence accuracy is coming lower. Train accuracy is quite high and hence data is overfitting. Data is becoming high dimensional compared to number of samples when we use tf-idf.

**Future Scope:** Use 20% train test split or use cross validation to get results and compare. Find a different loss function that deals with imbalanced classes. Apply more accurate augmentation for classes having a smaller number of samples. Apply feature reduction techniques so that better classification techniques can be used.