Ameliorating Questions Classification

By

Najam-ul-Sahar Arif

275496

Supervisor

Dr. Seemab Latif

Department of computing

A thesis submitted in partial fulfillment of the requirements for the degree of MS(IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(April, 2021)

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms. Najam-ul-Sahar Arif, (Registration No 00000275496), of MSIT-18 (School/College/Institute) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Supervisor: **Dr. Seemab Latif**

Date: _____

Signature(HOD): _____

Date: _____

Signature(Dean/Principal): _____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled '**Ameliorating Questions Classification**' submitted by **Najam-ul-Sahar Arif** have been found satisfactory for the requirement of degree.

Advisor: **Dr. Seemab Latif**

Signature:

Date:

Committee Member 1: **Dr. Sharifullah Khan**

Signature: _____

Date: _____

Committee Member 2: **Dr. Asad Ali Shah**

Signature: _____

Date: _____

Committee Member 3: **Dr. Qaiser Riaz**

Signature: _____

Date: _____

Abstract

One of the most vital steps in automatic Question Answering systems is question classification, also known as Answer type classification, identification or prediction. Precise and accurate question

classification can lead to the elimination of irrelevant candidate answers from the pool of answers available for the question. High accuracy of question classification means accurate answer for the given question. This paper proposes an approach, named as Question Sentence Embedding (QSE), for question classification by utilizing semantic features. Extracting large number of features do not solve the problem every time. Our proposed approach simplifies feature extraction stage by not extracting features such as named entities, present in fewer questions because of their short length, and hypernyms and hyponyms of a word that requires WordNet extension. These features make the system more dependent on external sources. We have used Universal Sentence Embedding with Transformer Encoder for obtaining sentence level embedding vector of fixed size and then calculated the semantic similarity among these vectors to classify questions in their predefined categories. As it is the time of global pandemic COVID-19 and people are more curios to ask questions about COVID-19. Our experimental dataset is publicly available COVID-Q dataset. Our results have achieved an accuracy of 69% on COVID-19 question classification task. Our proposed approach has outperformed the baseline method, 53.4%, manifesting the efficacy of proposed QSE method.

Keywords: COVID-19, Machine Learning, Multi-class, Question Answering Systems, Text Classification, Universal Sentence Encoder

Dedication

This thesis is dedicated to my beloved parents without whom this success would not be possible, also my brothers and my sister who support me through thick and thin and enable me to achieve this degree.

Acknowledgement

First and foremost, thanks to Almighty Allah and HIS Prophet (P.B.U.H) for bestowing me the knowledge and guiding me in pursuing MS degree. Constant support is required for accomplishment of anything. In this purpose, I would like to say special thanks to my Supervisor Dr. Seemab Latif for her continuous guidance and support. I would also like to thank my GEC members, Dr. Sharifullah Khan, Dr. Asad Ali Shah and Dr. Qaiser Riaz for giving advices to me in improving my work. I also like to say special thanks from deepest of my heart to my parents who are my constant support, who gives me confidence for achieving something great. Also thanks to my brothers and my sister for their support. I would also like to say thanks to my friends Aneeza khalid, Abeer Gauher, Tayyeba Riaz and Alia Umrani, without their help and constant support, I may not be able to cross this stage. In the end, I would like to thanks everyone who help me in any way at any time.

Contents

Contents

Chapter 1

1 Introduction

The basic introduction of the research has been covered in this chapter. This chapter starts by defining question answer systems. Then the different phases of question answer systems have been defined. Third phase of this chapter contains definition of question classification and steps included in question classification. The chapter is further proceeded by motivation, research questions, research objectives and problem statement. The chapter is ended by defining the structure of the thesis, other chapters of research included in this thesis.

1.1 Introduction

Question Answering is a rapidly growing research field that is catching researchers and user's attention towards itself [1]. Question answer systems are considered as advanced form of information retrieval and natural language processing. They are different from search engines in a way that they provide direct and accurate answers to the user instead of providing links and detailed answers that leads to wastage of time and sometimes users get fed-up of lengthy searching process. For the query like "What is the highest waterfall in United States?", instead of providing bundle of documents containing details about all the waterfalls in United States, the question answer system directly provides the name of only highest waterfall i.e. Olo'upena Falls [2].

From 1960's the research on question answering has begun [1] but still many question answer systems are not able to deliver accurate answers rather providing many links to the answer e.g.Intellexar.com. The early research was mainly focused on domain specific systems but in 2000's most of the research is occurring for generalized systems that satisfy all types of user's needs.

Some examples of question answer systems include MASQUE, FAQ FINDER, QAS precise and QUARC [3]. Based on the literature, the question answer systems can be classified into five further categories. These are forms of answers generated, data sources, domain of application, type of language and language paradigm [4]. In order to get the answer, the question passes through five

stages before providing accurate answer to the user. These stages are: Question classification, search engine, answer extraction, answer scoring and answer aggregation.

The general architecture of question answer system is as follows:

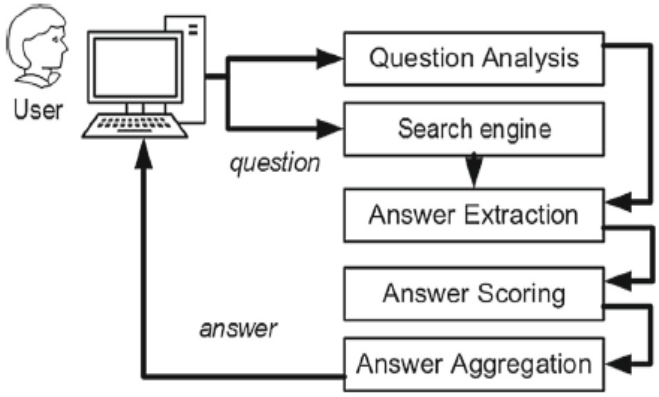

Fig 1.1: General architecture of QA systems taken from [5]

1.2 Question Classification

The first and most important step in automatic question answering is "Question classification". Question classification plays a vital role in achieving the accuracy of the question answer systems [2]. It has been already proved that almost 36.4% of the errors are happening due to misclassification of questions that lead to wrong answers and hence a collapsed system [2]. One of the most important ability of question classification phase is to reduce the search space that helps in increasing the efficiency of a system and reducing the time complexity.

The two main perspectives for question classification are:

1. Identification of surface patterns
2. Categorization of semantics

Surface example ID based methodology arranges the inquiries as the arrangement of word-based examples and answers are brought dependent on these examples. Such kind of question grouping technique experiences the limited ability to remove answers that have a place with not relevant classes. While as semantic classification based inquiry classifiers utilize outside information base like WordNet to characterize the inquiries dealing with hypernyms and synonyms.

Although text classification is a field similar to question classification with only a small difference that text classification ignores question WH-words such as What, When, Where, Who, why. These WH-words play an important role in question classification. The research in text classification is better done but question classification still needs some attention [2]. From the outset, one may feel that question classification can be encircled as a text classification task. Be that as it may, there exists qualities of question classification that recognize it from the regular errand. Initially, a question/ query is generally short and contains less word-based data than a lengthy document. Second, a short question needs a more profound level examination to uncover its shrouded semantics. Along these lines, application of text classification algorithms essentially to address classification problems couldn't result in a decent outcome.

Also natural language is very ambiguous field with many variations in text being written that conveys the same meaning. 'What' and 'Which' questions create more ambiguity than other questions. For example, for the question "What was the claim to fame of King Camp Gillette?" is of description type whereas the question "What is the name of highest point in world?" is of location type. Question classification is a very consequential field [6].

Below table shows how multiple questions convey same meaning and belong to same class.

Table 1.1: Multiple questions conveying same meaning

Question	**Class**
Will COVID stay and last perpetually?	Speculation
Will COVID ever go away?	Speculation
Will bequeath the COVID virus end?	Speculation

The question classification is also named as question analysis in some researches. It determines the expected answer/ Lexical answer type(LAT) of a question. Determining expected answer type helps to narrow down the search space and makes answer giving task easier for the question answer systems.

1.3 Motivation

The need for question answer systems is growing day-by-day. From open domain to specific domain, question answer systems are becoming need of the hour. Specific question answer systems

such as educational, Islamic and health care systems are becoming popular day-by day. Open domain question answer systems are a source of providing benefit to people and helps a lot to save time and give accurate response to users resulting in user comfort and popularity. Instead of using browsers, people now focus on using question answer systems. They do not need to browse through links and documents to acquire their required answer. The accuracy of these systems depend highly on the accuracy of question classification results. The question classifiers are not well accurate. They are either highly dependent on external sources. A domain specific question classifier is unable to achieve better results on open domain dataset and vice versa. So to build a generic and realistic model with more accurate results is very important. Also to study more features and extract the features that are of more importance than others is also needed. The question classifiers that are built to deal with specific domain questions such as medical domain or Islamic domain are utilizing such domain specific resources that they are useless for open domain questions. Building a generic and efficient model can help out in creating a question answer system that can be used in every domain with high accuracy results.

1.4 Problem Statement

Question Answer systems are becoming an important part of information systems. People are now giving more attention to Question Answer systems as compared to search engines. The reason for this is that Questions Answer systems are more reliable and time saving as compared to search engines where users have to invest more time for finding the relevant and accurate answer to their question.

There are many steps in building Question Answer systems such as Question classification, Answer extraction, Answer scoring and Answer aggregation. Question Classification is the first and most significant step in building up Question Answer systems. The purpose of question classifier is to assign one of the most accurate and relevant class or category among group of pre-defined categories or classes. If a question classifier is enough efficient to assign a question to the most relevant category and a question gets classified into correct category than chances for a Question Answer system to deliver accurate answer for the question to the end user highly increases.

In year 2019, a global pandemic known as COVID-19 or Coronavirus appears in the world and people are more curious to ask questions about its transmission, prevention, side effects, testing, origin and so more. There are hundreds of questions which were never asked before as well as questions which are unanswered until the date. People are asking questions on social websites, health websites and other sites such as Yahoo and CDC (Center for disease control). These questions asked by the public requires accurate answers as it the matter of health and life. There are no Question Answer systems built specifically for COVID-19 and so it is the need of time to build effective Question classifiers that classify COVID-19 questions into relevant classes with more accuracy.

1.5 Research Questions

COVID-19 Questions classification is an important factor to consider while building question answer systems that only answers COVID questions and helpful for health officials and public.
Some factors are important in this phenomena. One of which is to know about features that can be provide more accuracy in question classification process and factors which reduces time and human effort.
Motivation of the research encourages us to answer the following research
questions:

RQ1. What types of questions are included in the dataset?

RQ2. What features can be used for classification of questions in this domain?

RQ3. Which features provide better accuracy as compared to others?

1.6 Research Objectives

In order to answer the questions discussed in section 1.5, we have discussed here the objectives to highlights the steps needed to perform the research. Existing research have been studied in order to evaluate how different researchers studied Question answer systems, the classification of questions, dataset, features and methodology. A framework to classify questions in their pre-defined categories has been proposed. The purpose of this study was to evaluate COVID-19 dataset

and the features extracted to perform classification of COVID-19 questions. We have identified following research objectives in order to meet research questions.

RO1. To identify questions included in COVID-19 dataset.

RO2. To identify suitable features that can be extracted for classification purpose.

RO3. To analyze importance of feature extraction for classifying COVID-19 questions.

1.7 Thesis structure

The other chapters of the research will be in following structure. In literature review chapter 2 of this research, we will have examined literature from 1960 to 2020 that how researchers tried to achieve accuracy in question classification. What type of features they ensured to measure classify questions in classes. It will also identify the research gap. In chapter 3 methodology to implement question classifier for classifying questions will be discussed. System will be evaluated against multiple machine learning classifiers using stratified cross validation, various Classifier measures, different portions of dataset, classifier performance and classification evaluation will be discussed in chapter 4. At the end we will conclude our findings and future work in chapter 5 .

Chapter 2

Literature Review

Researchers have been working on question answering systems since late 1960's but proper research in this area has been started after the development of "START", world's first question answer system developed in 1993 [7]. The first taxonomy was given by Li and ROTH in [55]. After this huge progress started to appear in steps used in question answering systems and the most important of which is question classification.

2.1 Defining Question Classification

The formal definition of question classification [8] can be stated as:

Question classification is the task of assigning a boolean value to each pair hqj , cii ∈ Q × C, where Q is the domain of questions and C = {c1, c2, . . . , c|C|} is a set of predefined classes. Question Classification Assigning hqj , cii to the value T indicates that qj is judged to belong to the category ci , while an assignment to the value F indicates that qj is not judged as belonging to the category ci. In a machine learning setting, the task is to make the unknown target function Φˆ : Q × C → {T, F} approximate the ideal target function Φ : Q × C → {T, F}, such that Φˆ and Φ coincide as much as possible.

2.2 Taxonomies for Answer type

As discussed earlier, the main theme of question classification is to classify question in a set of predefined classes/categories [9]. A set of pre-defined classes is called a taxonomy. There are two types of taxonomies called as flat and hierarchical. Flat taxonomies are one level whereas hierarchical are multilevel. Flat taxonomies are not much effective as they put up all classes at the same level [10].

2.3 Question Types

Question types means the types or range of questions that can be asked [11]. There are many types described as follows:

Definition

This type of questions require answer in form of formal definition. Example: Define COVID-19.

Description

This type of questions must be answered in a detail definition. Example: Describe the Finnish music personality Salonen 's appearance.

Factoid

They require short form answer or a fact. Example: What is the name of current pandemic?

List

It requires the answer to be in the form of list. Example: What are the medicines used to treat COVID?

Procedural

Requires the response to be a rundown of guidelines for achieving a task. Example: How do I start a web based business?

Hypothetical

A hypothetical question depends on assumption, assessment, individual conviction, or guess, and not realities. Example: What would you do if you are given 24 hours to live?

Casual

It needs explanation of an entity. Example: Why does it snow?

Relationship

This type of questions demands answers as a relationship between entities or events, Example: How was Teddy Roosevelt related to FDR?

Opinion

Requires answer as an opinion about some event or entity. Example: What do you think who will be the president of this year's elections?

Confirmation

Answers in form of either YES or NO. Example: Are we playing a match tomorrow?

Table 2.1: Question types

Type of Questions	Function	Example
When	Mostly related to time	When COVID will be over?
What	Repetition/confirmation/information	What would happen if we not practice social distancing?
Where	Location	Where does the word COVID comes from ?
Who/Whom/Whose	Personality(Subject)	Who will find the COVID cure ?
Why	Reason	Why is quinine effective in curing COVID ?
How	Manner/ reason/condition	How COVID test is done?
Which	Related to choice	Which COVID antibody tests are precise?

2.4 Preprocessing

In field of textual data, preprocessing is the most important step. Without preprocessing text, applying machine learning algorithms on text are difficult to apply. Also the results would be totally different without preprocessing.

Stemming and lemmatization allows words to be in their root form [11].

2.5 Feature selection

There are four main approaches used for feature selection. The approaches are filter, wrapper, embedded and hybrid [12]. Feature selection improves the performance of machine learning classifiers.

[13] observed in their study that filter based feature selection approach is a faster approach but it does not guarantee accuracy. Filter based approach basically selects a small subset of features from a large dataset which is high dimensional. It does select features without the help of any leaning algorithm.

[14] stated in their study that wrapper based approach provides better feature selection accuracy for selection machine learning classifiers than filter method but its disadvantage is that it is very costly method. On the other hand, embedded method selects features during the phase of training and it also gives better accuracy for certain applied machine learning algorithms [15]. Hybrid method as explained by [16] is a combination of both a filter and a wrapper methods and provides accuracy better than individual approaches.

2.6 Feature Extraction

Extracting features from the dataset involves selecting original features from dataset that creates a combination of new ones [17]. Feature extraction is an important step in question classification. Three main feature types are Lexical, Syntactic and Semantic [17].

- Lexical Features

Features that represent context of the questions are known as lexical features [18] Lexical features includes Unigrams, Bigrams, N grams and interrogative pronouns. Unigrams are the features which are all single words in a question. They basically represent context of a question [19]. Bigrams are the combination of two consecutive words or terms [20]. Interrogative pronouns are the words that are important part of the question. The main difference between a sentence and a question is the interrogative pronoun [17] which are WH-words such as How, When, Where etc.

- Syntactic Features

Syntactic features are formal properties of syntactic items which decide how they carry on concerning syntactic limitations and tasks (like determination, permitting, understanding, and development). The two main syntactic features are Parts of speech tags and head words [20]. Parts of speech includes noun, pronoun, adverb, adjective and so on. POS tags play an important role in classification of questions. Heads words are the most important features. They can be one of the POS tag present in the question or sometimes it can be a bigram or trigram.

- Semantic Features

Semantics represent the meanings of the question. They include synonyms, hypernyms, hyponyms and related words in a question [21]. These features are extracted using WordNet [21] observed in their study that WordNet based semantic feature extraction improves classification accuracy. Synonyms are the meanings of a words. Hypernyms are the broad categories of a word. They represent broader concept such as Color is a broader meaning of purple, red, orange, yellow. While

Authors in [22] used dependency trees as a set of semantic features.

2.7 Types of Classification

A question is classified into one or many categories from pre-defined taxonomy. Based on this, three types of question classification can be done.

2.7.1 Binary Classification

If the number of categories in the taxonomy are two and question is assigned precisely to one of the two categories, it is said to be as binary classification [23]. For example, in news detection, a news can be either fake or real.

2.7.2 Single Label Classification

In a system where a taxonomy consists of many classes/ categories but a question is assigned to exactly one category only, this is known as single label classification [24]. Example, In COVID-

19 questions classification, either a question can be asked about transmission or protection but a single question does not belong to both of these categories.

2.7.3 Multi Label Classification

If a taxonomy consists of multiple categories and each question or text belongs to more than one category than it is known as multi label classification [25]. Multi label classification is further divided into flat and hierarchical classification [26]. In flat classification, no relationship between categories exist whereas in hierarchical there each category has a sub category known as fine grain and coarse grain classes.

Chapter 3

3. Methodology

In this section, methodology for classifying COVID-19 questions is discussed in detail. The framework for classifying questions classifies questions in one of the 15 classes and helps in developing more efficient Question Answer systems. First of all, we have studied the dataset provided by Wei et al. [27]. The baseline dataset contains questions related to the global pandemic Coronavirus also known as COVID-19. The taxonomy consists of fifteen classes / categories and questions in the dataset classified into one of these classes. The questions are first pre-processed then passed to the feed forward transformer encoder module of universal sentence encoder. Then a similarity matrix is formed with these 512 fixed size vectors. Feature extraction is the most important stage in any system. In our system we do not extract features which are less contributing towards the accuracy of system like named entities. Questions which are more similar are placed into the same class. The main advantage of question classification is that it helps in fetching accurate answers for the user asked questions in a question answer system [28].

The underlying section is divided into two sub-sections i.e. system design and evaluation settings. Proposed system will be evaluated using quantitative approach and in the "evaluation settings" part all the measures used to evaluate the system will be discussed. The architecture is shown in fig 3.1.

3.1 System Design

The architecture of the proposed system consists of five steps:

(1) Preprocessing

(2) Feature Extraction

(3) Semantic Similarity

(4) Machine Learning Classifier

(5) Question Classification

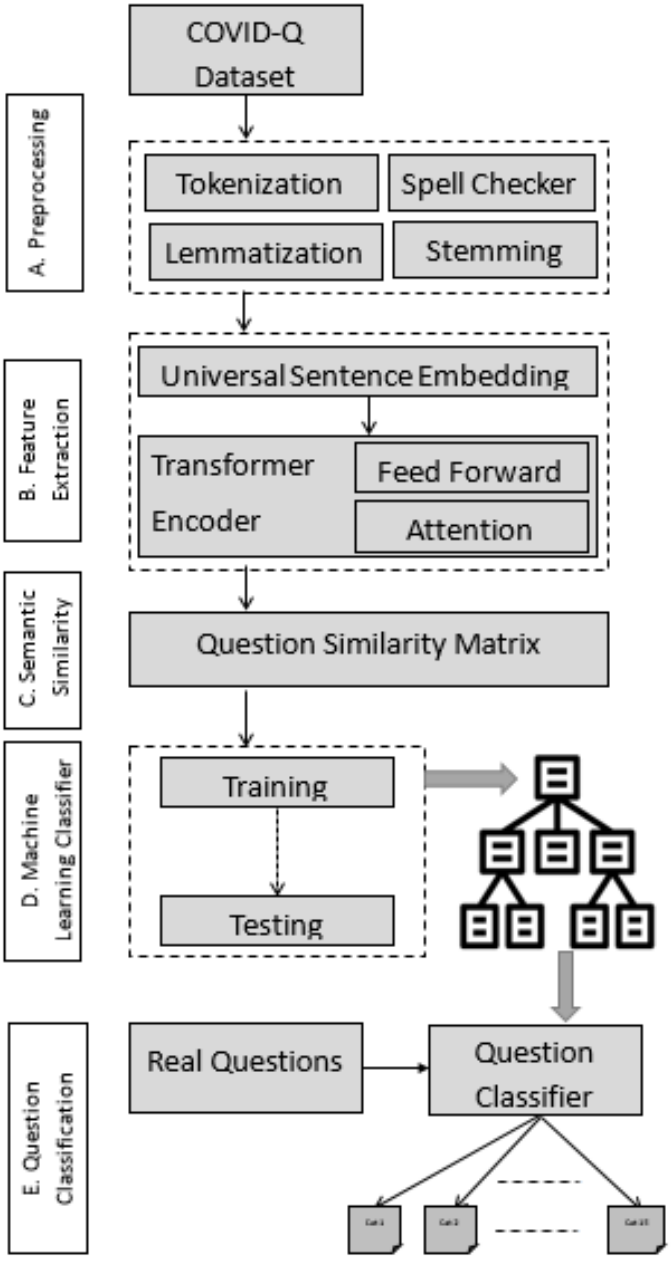

Fig 3.1: Proposed Architecture

3.1.1 Preprocessing

One of the fundamental steps in NLP tasks is to preprocess the textual data to removes irregularity, ineffectualness and noise from the data and convert it into an intelligent structure [29]. In preprocessing, we have first tokenized the questions into words.

In case of questions asked by users, there are high chances of misspelled words. Training a system on misspelled words may lead to errors and poor results [30,56]. For this reason, we have applied a spell checker module which checks each word in the dataset and corrects misspelled words converting data in a more appropriate form. After this, lemmatization and stemming was performed. We specifically do not remove stop words because the length of the questions is short and removal of stop words may lead to the removal of actual context of the question. For example, in the question "Will COVID go away in spring?" all of the words except COVID and spring are stop words. Removing them will remove whole question context and hence decreasing the accuracy of the classifier. The dataset is then normalized to be made available for the next feature extraction module.

3.1.2 Tokenization

In this stage, each word separated by a space is tokenized. Each paragraph or sentence is divided into terms which are called as tokens. Certain conditions for tokenization can be used. For instance, tokens can be produced on the basis of space, on the basis of dot or comma or some regular expression [31]. It is the essential stage to carry out advance preprocessing steps on information like extraordinary character evacuation. As information could contain crude qualities like alphabetic, numeric and both alphanumeric, whitespaces so it was difficult to carry out this step. White space tokenizer was used for this purpose.

3.1.3 Spell Checker

This is one of the essential step in natural language processing. As data is raw and spelling mistakes are common in raw data such as 'COVID' can be mistakenly written as 'COVED' or 'antibody' can be written as 'antebody' etc. As the length of a question is small and each word has its significance. So if a word is not correctly spelled, it would create other errors in the feature extraction module. Therefore, spell checker module performs an important task.

3.1.4 Stemming and Lemmatization

Stemming and lemmatization are important steps for data preprocessing [32]. After tokenization and correction of misspelled words, the words or terms needs to be stem through some good quality

stemmers. Stemming and lemmatization chop the word to its root form making it easy to process data. It reduces the length of data. Remove overhead of data handling. For stemming, porter stemmer was used. Porter stemmer is the large margin and most widely used stemmer with proficiency over others. For lemmatization, WordNet Lemmatizer was used which is itself a huge library of words and performs lemmatization more accurately.

3.1.5 Stop words removal

Stop words are the words that are of no special use and they do not deliver any meaning. They are only present to support verbs, objects and subjects in a sentence. Some examples of stop words are 'is, am are, also, in, and, the' [33]. The list of stop words in English language is not small. Removing stop words reduces the size of data as well as makes data handling easy. As length of the questions is small and each word is important so we didn't remove stop words from COVID-19 questions.

3.2 Feature Extraction

In this step, the output from previous stage is fed into the Transformer Encoder module. Universal Sentence Encoder (USE) is used to encode the input before it is passed on to transformer encoder module, see Fig. 3.2. The transformer encoder is based on original transformer architecture [30]. This architecture consists of six layers of transformers stacked onto each other. The two main are multi headed self attention and feed forward network which converts text into fixed sized vector. The layers are based on feed forward network. Each layer also consists of a self-attention module that generates the representation of words by taking into account the order of the words as well as the surrounding texts that delivers the context of each question. The output generated by this

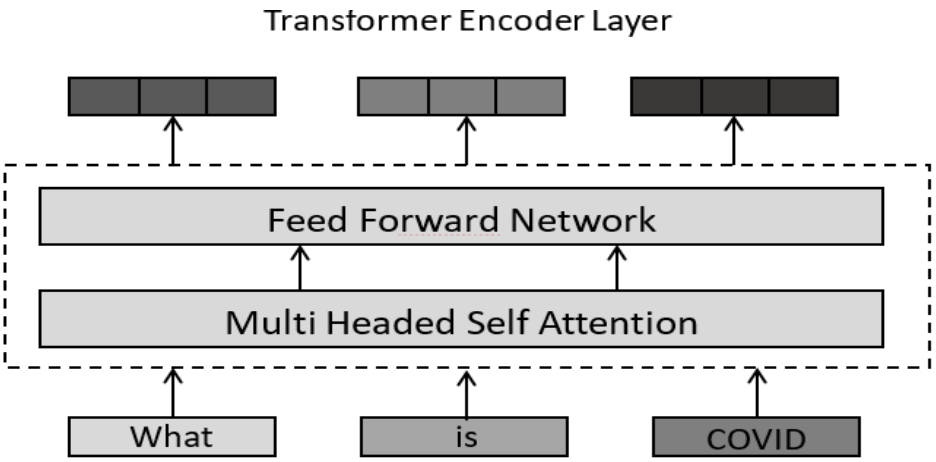

Fig 3.2: Transformer architecture

module are feature vectors that are context aware word embedding. These context aware word embedding are placed element wise and then a module divide them with the square root of the length of each sentence, see Fig 3.3.

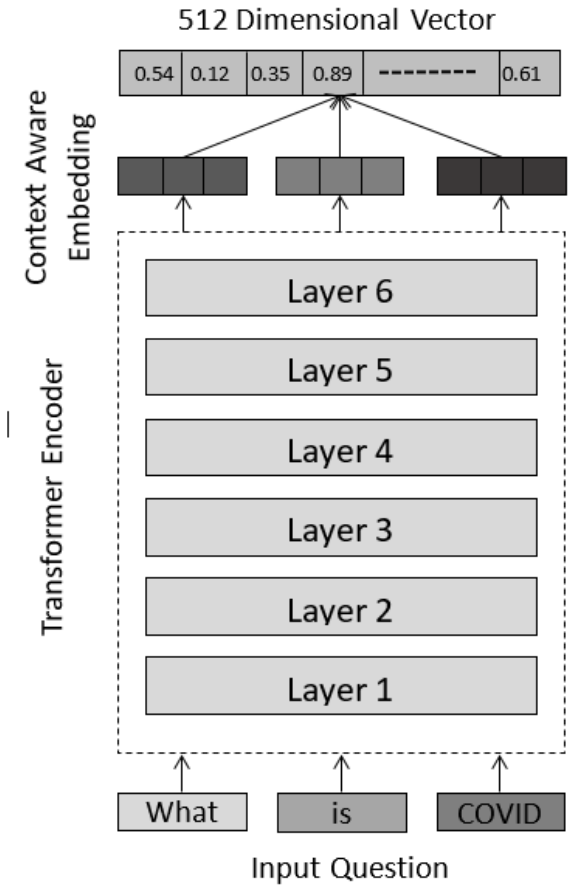

Fig 3.3: Universal sentence encoder

This step is done in order to deal with the problem of different lengths of questions that are asked by the user. This will result in fixed size 512 dimensional vectors. Apart from the questions, the categories of the questions are also labeled by implementing the label encoder module. The major advantage of USE over traditional embedding such as Glove and word2vec is that it computes the embedding by taking whole sentence into account and generates 512 sized dense vectors. Traditional methods compute sentence embedding by calculating average of each word in the sentence. These averages then represent the whole sentence. The process of average calculation results in loss of information present in the sentence as well as it does not take into account the order of words in the sentence. Whereas, in terms of short sentences such as questions, the order matters a lot and change in order results in loss of context as well.

	text	label
0	will covid ever go away	Speculation
1	was covid predicted	Speculation
2	when covid will be over	Speculation
3	will covid be gone this year	Speculation
4	when will the covid virus end	Speculation

Fig 3.5: Questions with labels from COVID-19 Dataset

	text_Embeddings	label	label_encode
0	[0.07343219965696335, -0.023441066965460777, -...	Speculation	10
1	[-0.020498821511864662, -0.05966268479824066, ...	Speculation	10
2	[0.018561743199825287, -0.03658769652247429, -...	Speculation	10
3	[-0.004635238088667393, -0.10405649244785309, ...	Speculation	10
4	[0.05261843651533127, -0.06817462295293808, -0...	Speculation	10

Fig 3.6: Questions labels with encoding

For example, “Will COVID go away in spring” and “Will spring go away in COVID” are two different questions with totally different answers and context but traditional embedding methods generate the same output while USE generates the separate embedding vectors by keeping the context of questions in account. The USE is pre-trained on millions of data from multiple sources including Wikipedia [34]. It augments unsupervised learning with training on supervised data corpus such as Stanford Natural Language Inference (SNLI) for improving performance of USE. Fig 3.3 shows the conversion of questions from the dataset into 512 dimensional vectors with encoded labels.

3.3 Semantic Similarity

In this step, we find the questions that are semantically similar and assign them categories. Cosine similarity is one of the powerful measure to find similarity between two vectors by calculating the cosine of the angle between them using equation given in 1. For two similar vectors, cosine function

is 1 when theta = 0, while for two non-similar vectors, cosine function is -1 and theta=180 [35]. The problem in our case is the limited amount of dataset. For this reason, we used the embedded feature vectors and calculate semantic similarity among the questions by simply calculating dot

product of these vectors and setting a threshold limit of 65% to get more similar questions that belong to the same category. These two feature vectors are then fed into a machine learning classifier. The best results are acquired when these features are fed into SVM polynomial kernel.

$$\text{Cos}(\theta) = \frac{A.B}{||A||.||B||}$$

$$= \frac{\sum A_i B_i}{\sum Ai^2 \sum Bi^2} \quad (1)$$

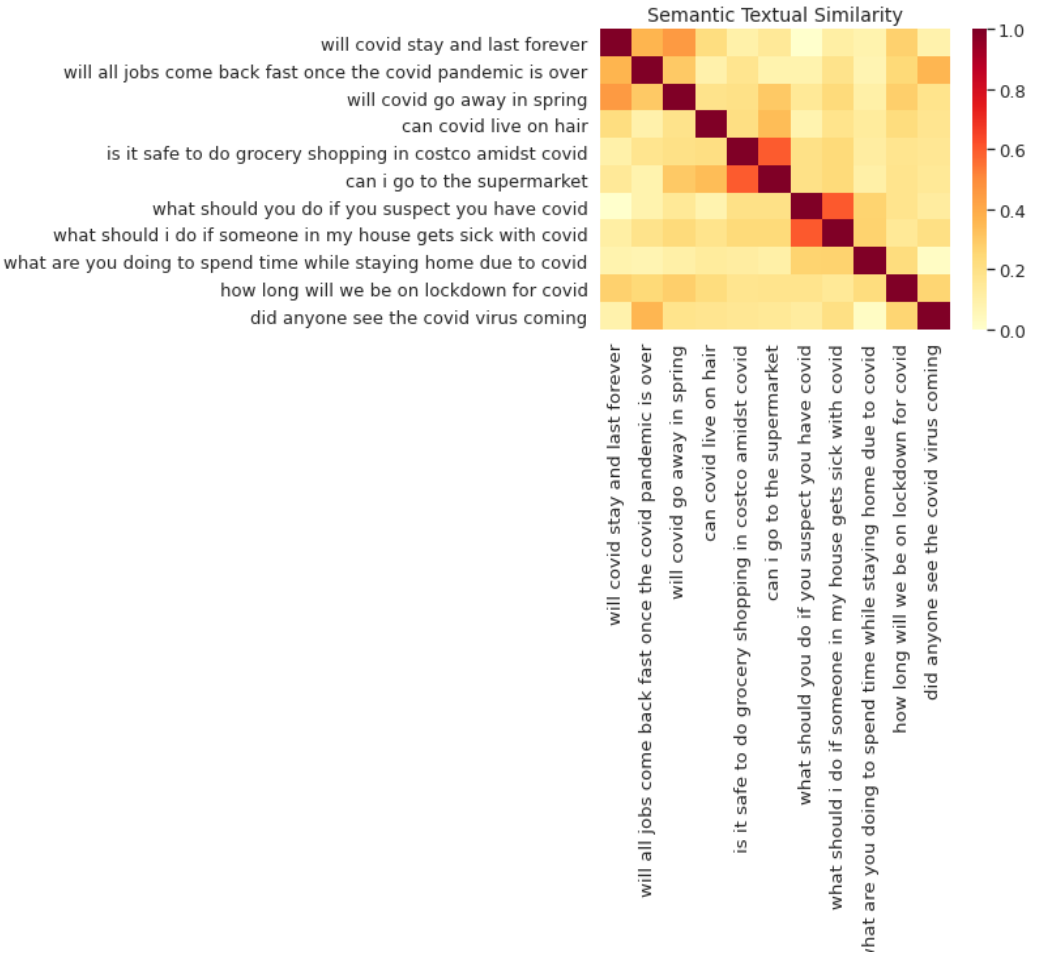

Fig 3.7: Semantic similarity among questions

Machine Learning Classifier

We train our machine on five different classifiers which are Stochastic Gradient Descent, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors and Decision trees. We train machine on augmented as well as non-augmented datasets and test on three different test sets named as Real Q, Generated Q and third one is the combination of both real and generated datasets. We used stratified k fold to solve over-fitting and under-fitting problems.

3.4 Evaluation Settings

In order to evaluate the proposed methodology, the selected dataset and evaluation metrics are discussed in detail below.

3.4.1 Dataset

We have studied the baseline dataset provided by Wei et al. [27]. The data was gathered from 13 different sources such as CDC, Food and Drugs Association (FDA), John Hopkins University and other crowd sourced sites like Yahoo, Bing and Quora. The dataset contains questions asked by people related to COVID such as “When will COVID end?” and “Who is at a higher risk for serious illness from COVID?”.

Table 3.1: Data split details

Question Categories	15
Training Questions per Category	20
Training Questions	300
Test Questions (Real)	668
Test Questions (Generated)	238

The dataset consists of 1690 unique questions divided into 15 categories based on the type of question. Question categories are given in Fig 3.8. Most of the questions asked in the dataset are about prevention from COVID-19, societal effects, and transmission of COVID-19. The questions are factoid, list, definition, description, opinion, casual, relationship and procedural. Example of these types of questions from the dataset are given in Table 3.2. The questions in the dataset are manually annotated and validated again by external annotators. There were three rounds of manual annotations that made question categories more reliable.

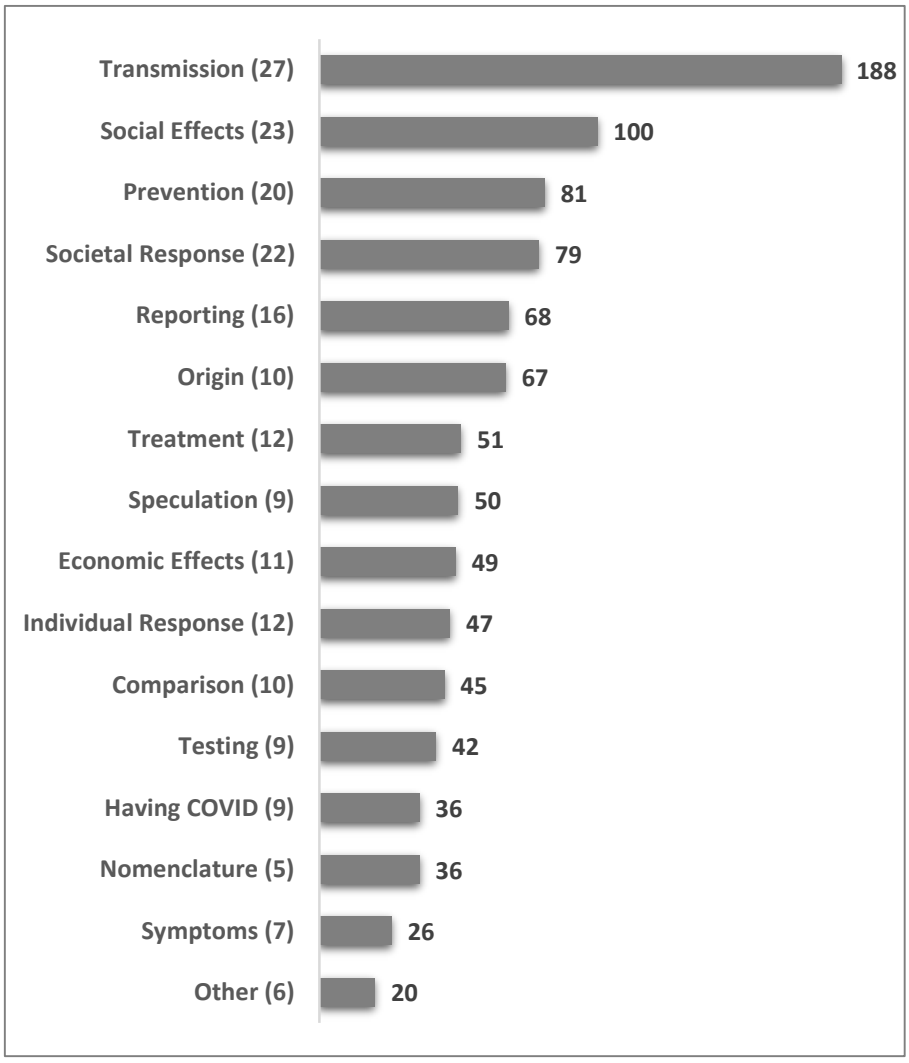

Fig 3.8: Dataset categories

Table 3.2: Question types

Factoid	From where does COVID originated?
List	What are the early symptoms of COVID-19?
Definition	What does COVID means?
Description	Why are people more concerned with going back to work than staying home until the COVID virus improves?
Opinion	What would happen if TRUMP tests positive for COVID-19?
Casual	Why is it being advised to not take ibuprofen if one has COVID?
Relationship	What is the difference between COVID and SARS Cov 2?
Procedural	how does the COVID virus cause death in the infected patient?

3.2.2 Evaluation Metrics

The evaluation metrics used to appraise the performance of our approach and baseline approach was accuracy. However, we additionally provide results with precision, recall and F1 measure. Accuracy is the metric used to calculate the effectiveness of classification models. It is the ratio of questions classified correctly by the model over the total number of questions classified in their predefined categories by the model as shown in equation 8. To measure quality, precision metric is used. Precision is a measure of true positives (the questions accurately categorized by the model) predicted by the model divided by the total number of true positives and true negatives predicted by the model as shown in equation 2. Recall is basically the quantitative measure. It is the ratio of questions belong to the positive category and correctly classified by the model by the total number of questions that actually belong to the positive categories as shown in equation 3. F1 measure commonly known as F1 score is a balanced measure which takes precision as well as recall into account.

3.2.2.3 Micro-average Measures

In order to evaluate the performance of model, three basic micro average measures have been used:

Table 3.3: System Actual and predicted values

Actual	**Predicted**	
	Yes	**No**
Yes	TP	FN
No	FP	TN

- Micro-averaged precision: It is the sum of true positives divided by the sum of true positives and true negatives [36]. Formula is given in (2).

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

- Micro-averaged recall: It is the sum of true positives divided by the sum of true positives and false negatives [37]. The formula is shown in (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

- Micro- averaged F1 score: It is the harmonic mean function of precision and recall that is needed to provide balance between precision and recall [38]. The formula of F1 is given in (4).

$$F1\ Measure = 2 \times \frac{Precision\ \times Recall}{Precision + Recall} \quad (4)$$

3.2.2.3 Macro-average Measures

In order to evaluate the performance of model, along with micro average measures, macro average measures have been used. These are explaining below:

- Macro-averaged precision: This measure is used to evaluate system's performance on all classes altogether. Formula to calculate is given in (5).

$$Precision = \frac{1}{N}\sum^{N}\frac{TP}{TP + FP} \quad (5)$$

- Macro-averaged recall: It is the average of recall of system on all classes/ categories. The formula to calculate this measure is shown in (6).

$$Recall = \frac{1}{N}\sum^{N}\frac{TP}{TP + FN} \quad (6)$$

- Macro-averaged F1 score: It is the weighted harmonic mean of macro averaged precision and recall measures. The formula is shown in (7).

$$F1\ Measure = 2 \times \frac{Precision\ \times Recall}{Precision + Recall} \quad (7)$$

3.2.2.3 Weighted average

This measure is used in case when in dataset some points are given more importance as compared to others. It is calculated by calculating average of precision, recall and F1 measure of each class and then multiply the value by the weight of each point [39].

3.2.2.4 Accuracy

It is the measure of extent by which result of certain calculation or measurement conforms that it is the correct value [40]. The formula for this measure is given below.

$$Accuracy = \frac{\text{Number of correctly classified questions}}{\text{Total number of classified questions}} \quad (8)$$

Chapter 4

4. Evaluation and Discussion

In this section of the thesis, we will evaluate the performance of the system on COVID-19 dataset. As COVID has already disturbed whole world and appears as a global pandemic so it is very important to evaluate COVID data and provide results that may help in any way [51],[52],[53],[54]. Also many datasets have been made such as COVID twitter dataset, COVID Facebook data and a dataset consists of COVID research papers [50]. In order to evaluate the performance of the algorithms, we have trained our system on two datasets named as augmented and non-augmented while test them on three different datasets named as real dataset, generated dataset and third dataset is a combination of both real and generated.

We evaluate the performance of system through four different machine learning classifiers Stochastic gradient descent (SGD), Support vector machine (SVM), K-nearest neighbor (KNN), Decision trees (DT). Additionally, we extract some other features to check their impact on results as well.

4.1 Classifier performance

We have followed the same data split configurations as used by Wei et al. [27]. Table 3.1 shows the data split for question category classification. In our first experiment, we choose universal sentence embedding based transformer architecture to obtain sentence level embedding of each question. Then we calculated the semantic similarity among these obtained questions embedding. Feeding these features into the machine learning classifier SVM with polynomial kernel has achieved accuracy better than the baseline approach. Training on the non-augmented data, the baseline approach accuracy was 53.4% while our approach achieved 69% accuracy.

We have also calculated the accuracy on both test sets, Real and Generated, to evaluate the effectiveness of our approach. The achieved accuracy on Real+ Generated original test is 63% while on augmented Real+ Generated datasets accuracy fell down to 60%. Comparative results are given in table. In the second experiment, we fed the features into four different machine learning

classifiers and proved that our proposed QSE approach has outperformed all other approaches. First we fed features into Stochastic Gradient Descent (SDG) that randomly chooses a function point and then finds the minimum point in the function by stepping down its slope.

The SDG provides an accuracy of 63%, results are given on Table 4.4. Then we fed features into K-Nearest Neighbor (KNN) and Decision Tree (DT) classifiers. We choose two nearest neighbors in KNN and max depth of 2000 leaf nodes in case of DT. In case of KNN and DT, the accuracy drops down noticeably. This is because KNN depends largely on quality of data and scales linearly to the amount of data. As our dataset is not large, so KNN's performance is not good. Decision tress also do not perform well when number of categories are large that is the case with COVID-Q dataset and DT also have higher probability of over fitting. They give lower accuracy as compared to other machine learning classifiers. In the third experiment, we have extracted Positive Pointwise Mutual Information (PPMI), which are the association of correlation between a question word and a category. PPMI values as a feature declined the accuracy when fed into SGD, NB and SVM classifiers.

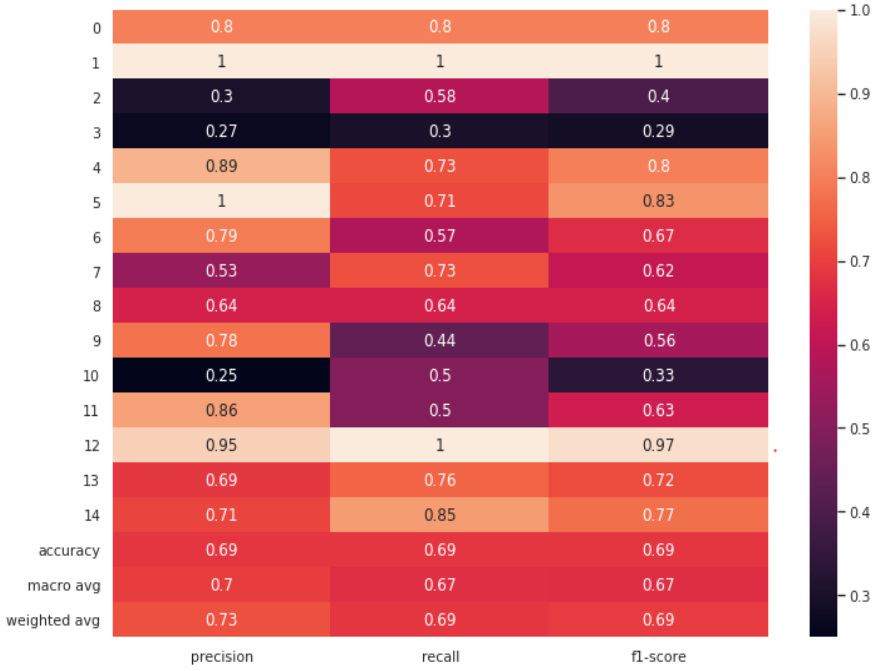

Fig 4.1: Accuracy of Classifier

There are 15 categories in COVID-19 dataset. Each question belongs to one category. There is an unbalance in train and test data. In train data, nomenclature is the class that contains least number of questions whereas in test data speculation is the class with least number of questions. This unbalancing in dataset is shown in fig 4.2.

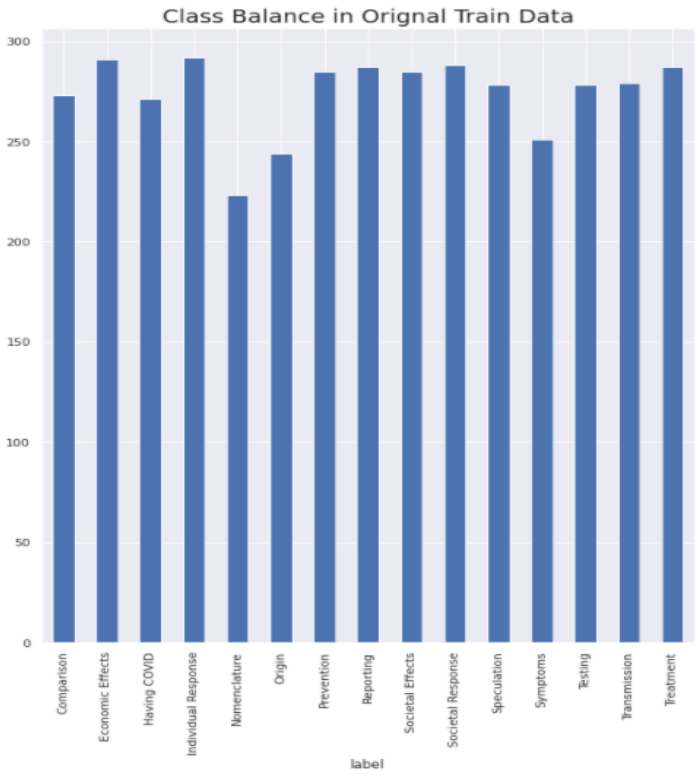

Fig 4.2: Class Balance in train data

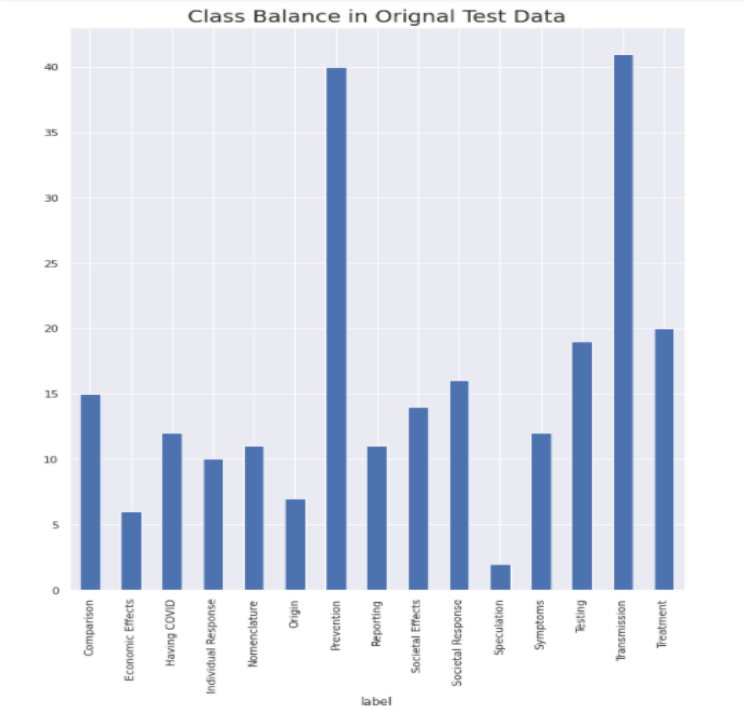

Fig 4.3: Class balance in test data

4.3 Impact of BERT VS. USE on COVID-19 dataset

In the table given below, the classification results of baseline methodology and our methodology are stated for real as well as generated datasets. The baseline methodology involves BERT for the purpose of feature extraction therefore in table 4.1, it can be seen clearly that performance of BERT on COVID-19 dataset is not good. It is giving average results only. On the other hand, transformer encoder of universal sentence encoder improves the classification performance very much on COVID-19 dataset. The main reason for this difference of performance is that BERT is trained on few tasks such as next sentence prediction and to predict the missing words in a sentence and not specifically trained on classification problems [41]. While Universal sentence encoder is trained on number of sentences where main task is of classification.

Table 4.1: Result comparisons of base and proposed methodology

Model	Real Q	Generated Q	Real + Generated Q
BERT+KNN	47.8	52.1	-
Augmentation	47.3	52.5	-
BERT+SVM	52.2	53.4	-
Augmentation	58.1	58.8	-
QSE(Our approach)	61.0	69.0	63.0
Augmentation	58	66.0	60.0

4.4 Classification with Positive pointwise mutual information(PPMI)

In this section, we'll explain the impact of extracting positive pointwise mutual information from the dataset and fed it into the machine learning classifier as a feature set. Pointwise mutual information is basically an alternative of raw frequency. It is the ratio of co-occurrence of two words together [42]. It measures when a word let's say 'a' occurs with another context word say 'b' and both of them together gives a meaning. Pointwise mutual information results in positive and negative values. When two words are co-related, it gives positive values and for non-related words, it gives negative values. The positive values are known 'Positive pointwise mutual information'.

The formula for PMI and PPMI are written below;

$$PMI(y,z) = \log\left(\frac{P(y,z)}{P(y)P(z)}\right) \tag{9}$$

$$PPMI(y,z) = \max\left(\log_2 \frac{P(y,z)}{P(y)\,P(z)}, 0\right) \tag{10}$$

PMI is positive either in one of these three cases:

1- When $\frac{P(y,z)}{P(y)P(z)} > 0$

2- When $P(y,z) > P(y)P(z)$

3- When two words occurring together are more meaningful then occurring individually. For Example, kick and Ball occurring together are better than individual.

In our case, when we extract PPMI as one additional feature, it drops accuracy to much low level which proves that PPMI is not a good feature in this case to extract. This feature has negative impacts on accuracy of COVID-19 questions classification. The effects of PPMI on dataset and accuracy are shown in table 4.2 below.

Table 4.2: Results with PPMI

Train	Test	Classifiers		
		SGD	**SVM**	**NB**
Non-Augmented	Real	0.41	0.37	0.31
Non-Augmented	Generated	0.40	0.31	0.31
Non-Augmented	Real+ Generated	0.45	0.40	0.39

4.5 Classification Evaluation

We have stated results of four different classifiers. When output of universal sentence embedding and similarity matrix entered into machine learning classifiers, they generate accuracy. In our first experiment, we fed features into Stochastic gradient descent (SGD) classifier. SGD randomly chooses a function point and then finds the minimum point in the function by stepping down its slope. SGD is a basic building block of neural networks. In cases where optimal points in a function are unknown, SGD equates the slope of function to 0 [43], [44].

In our second experiment, we choose Decision trees as a machine learning classifier to fed features into it and get results. Decision trees are a tree like structure where there are internal nodes, leafs and edges.

A DT is a prescient model communicated as a recursive segment of the element space to subspaces that comprise a reason for forecast. A DT is established coordinated tree. In DTs, hubs with active edges are the interior hubs [45], [46]. Any remaining hubs are terminal hubs or leaves of the DT. DTs order utilizing a bunch of progressive choices on the highlights. The choices made at inner hubs are the part rule. In DTs, each leaf is appointed to one class or its likelihood. Little varieties in the preparation set outcomes in various parts prompting an alternate DT. Consequently, the mistake commitment because of difference is enormous for DTs. One of the noticeable advantage of decision trees is that they select features in a variable manner instead of static selection of variables. But on the other side, the biggest advantage includes over-fitting, biasness and variance that is caused by just a small change in data. The highest accuracy given by DT on COVID-19 dataset is 37% on real+ generated questions.

In third experiment, we fed features into K-nearest neighbor that is a machine learning classifier. KNN is considered a good classifier for classification and regression problems. KNN is one of the simplest algorithm for classification problems [47]. It is based on the idea of selection of similar data points in a cluster. We choose the value of K=2 which is the number of nearest neighbors that are similar to each other. In step 1. We choose random data point from COVID-19 dataset then calculate distance between nearest neighbors. This forms the clusters of similar data values.

In this case, the noticeable disadvantage of KNN is that its speed slow down with increase number of data points. In real scenario where number of data points are large, KNN's performance slows down. In our case of COVID-19 dataset, as it is not very large dataset, the performance of algorithm is pretty fine acceptable. It gives 53% on real and straight increase of 5% reaching the accuracy to 58% in case of generated questions.

In fourth experiment, we fed features into Support vector machines (SVM). SVM is based on hyperplanes which is the maximum margin/distance between two data points separating two classes. The points on which hyperplanes are based are called as support vectors. If support vectors are change then hyperplane also changes [48], [27], [49]. We fed features into SVM, which gives maximum and best accuracy as compared to other machine learning classifiers and much better as compared to base paper. The accuracy is 69% which is considerably noticeable and gives weight to our proposed methodology.

Table 4.4: Results with multiple classifiers

Classifiers	Real Q	Generated Q	Real Q + Generated Q
USE + SGD	56.1	63.0	58.0
Augmentation	56.0	62.0	58.0
USE + DT	35.0	33.0	30.0
Augmentation	35.0	39.0	37.0
USE + KNN	53.0	58.0	55.0
Augmentation	54.0	56.0	55.1
QSE (USE + SVM)	61.0	69.0	63.0
Augmentation	58.0	66.0	60.0

In order to make results clear, accuracy comparison for different machine learning classifiers are shown in fig 4.4.

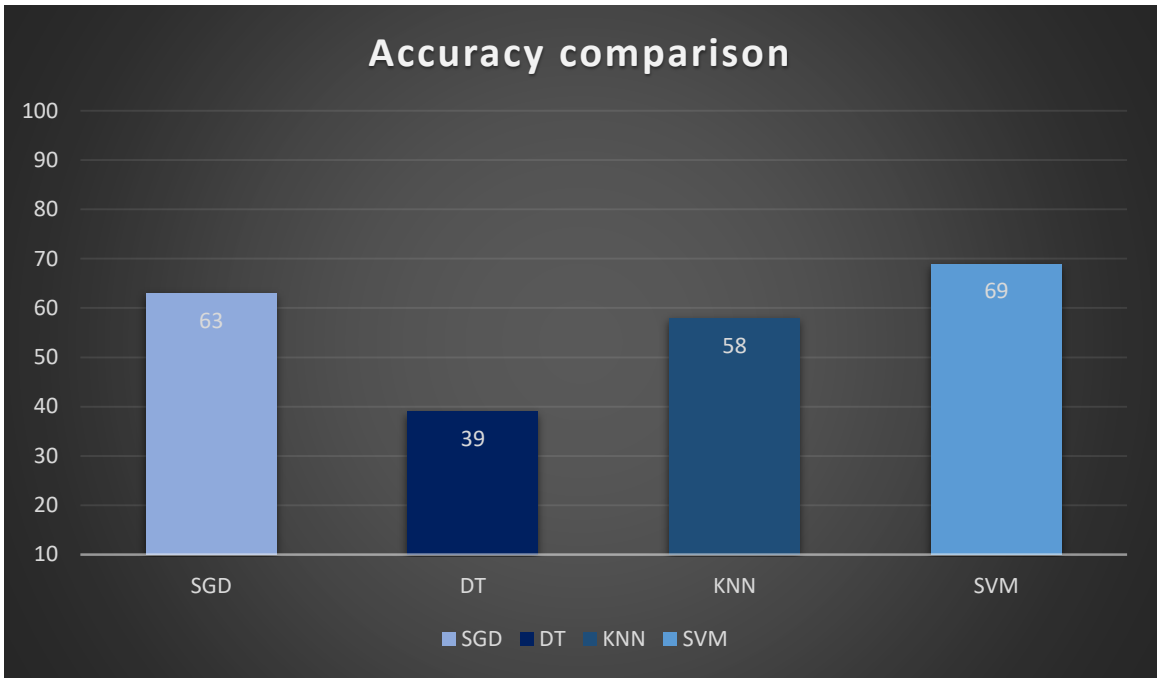

Fig 4.4: Results with different ML classifiers

4.6 Comparison between base methodology and QSE

In work presented by Wei et al [27], their methodology is based on bidirectional encoder representations from transformers with SVM and KNN machine learning classifiers. The main disadvantage of BERT is its cost at big scale. Also the main purpose to build BERT was to predict next sentence or to predict missing words in one sentence. Our problem is to classify questions that helps to build proficient question answer systems.

Our methodology is based on Universal sentence embedding with transformer architecture. The universal sentence encoder in comparison to BERT is trained on sentences and its main objective is to find similarity between sentences. Also Universal sentence embedder is built up on unsupervised training data from web Question answer pages and discussion forums. This training on unsupervised data make its use effective for Question answer systems such as COVID-19 questions dataset.

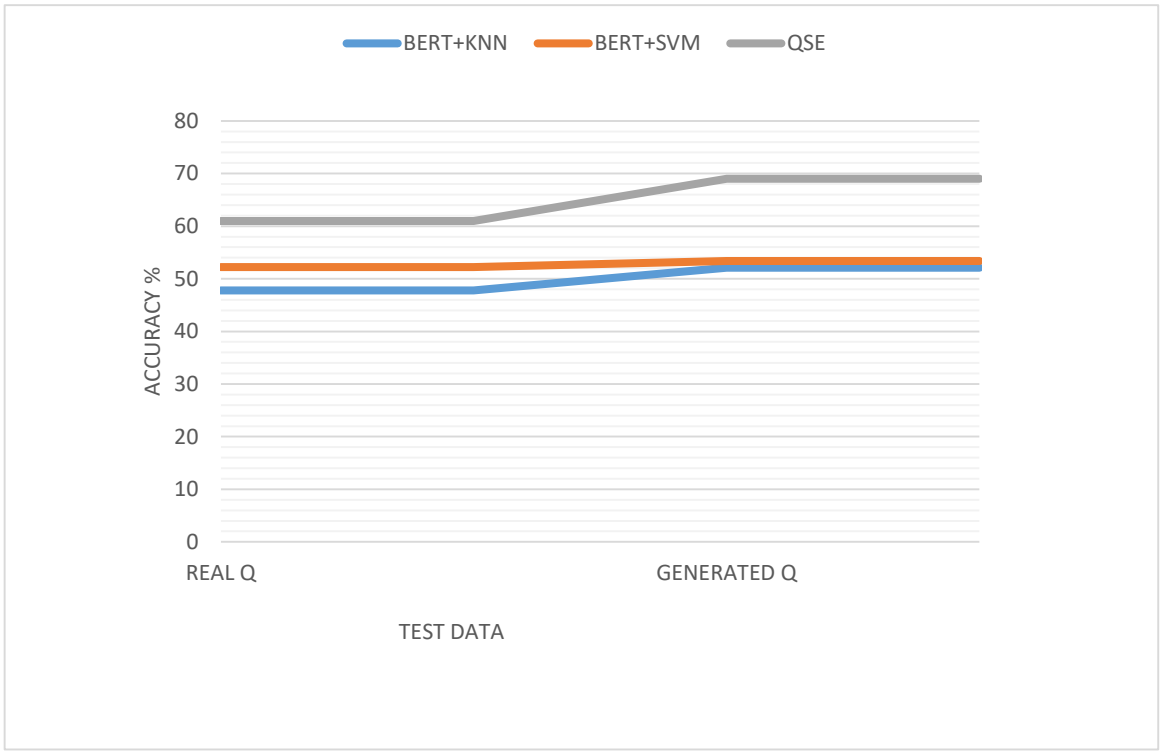

Fig 4.5: Accuracy comparison

Chapter 5

5 Conclusion

Aiming for classifying questions into predefined categories for the development of COVID question answer systems, we design an approach named as QSE. This approach is tested on COVID-Q dataset which consists of Corona virus related publicly asked questions. Our approach uses Universal Sentence Encoder based Transformer architecture. We fed compact set of features overcoming the limitations of lexical and syntactic features and named entities which do not work well in short questions. “What does COVID stands for?” do not contain any named entities and hence acquiring these features result in complexity. By acquiring universal sentence embedding for COVID-Q questions along with semantic similarity, our approach achieves significant higher accuracy than the baseline accuracy reported by the authors.

References

[1] Soares, M. A. C., & Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, *32*(6), 635-646

[2] Mohd, M., & Hashmy, R. (2018). Question classification using a knowledge-based semantic kernel. In *Soft Computing: Theories and Applications* (pp. 599-606). Springer, Singapore.

[3] Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University-Computer and Information Sciences*, *28*(3), 345-361.

[4] Ojokoh, B., & Adebisi, E. (2018). A review of question answering systems. *Journal of Web Engineering*, *17*(8), 717-758

[5] A. A. Shah, S. D. Ravana, S. Hamid, and M. A. Ismail, "Web pages credibility scores for improving accuracy of answers in web-based question answering systems," IEEE Access, vol. 8, pp. 141456–141471, 2020.

[6] P. Le-Hong, X.-H. Phan, and T.-D. Nguyen, "Using dependency analysis to improve question classification," in Knowledge and Systems Engineering (V.-H. Nguyen, A.-C. Le, and V.-N. Huynh, eds.), (Cham), pp. 653–665, Springer International Publishing, 2015.

[7] D. Zhang and W. S. Lee, "Question classification using support vector machines," in Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, (New York, NY, USA), p. 26–32, Association for Computing Machinery, 2003.

[8] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

[9] H. Hardy and Y. Cheah, "Question classification using extreme learning machine on semantic features," Journal of ICT Research and Applications, vol. 7, pp. 36–58, 2013.

[10] Z. Huang, M. Thint, and Z. Qin, "Question classification using head words and their hypernyms," in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, (Honolulu, Hawaii), pp. 927–936, Association for Computational Linguistics, Oct. 2008.

[11] Hao, T., Xie, W., Wu, Q., Weng, H., & Qu, Y. (2017). Leveraging question target word features through semantic relation expansion for answer type classification. *Knowledge-Based Systems*, *133*, 43-52.

[12] Hoque, N., Singh, M., & Bhattacharyya, D. K. (2018). EFS-MI: an ensemble feature selection method for classification. *Complex & Intelligent Systems*, *4*(2), 105-118.

[13] Kumar, S., Garg, S., Mehta, K., & Rasiwasia, N. (2019, November). Improving answer selection and answer triggering using hard negatives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5911-5917).

[14] Yu, M., Guo, X., Yi, J., Chang, S., Potdar, S., Cheng, Y., ... & Zhou, B. (2018). Diverse few-shot text classification with multiple metrics. *arXiv preprint arXiv:1805.07513*.

[15]] Iyyer, M., Manjunatha, V., Boyd-Graber, J., & Daumé III, H. (2015, July). Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1681-1691).

[16] Chen, Q., Peng, Y., & Lu, Z. (2019, June). BioSentVec: creating sentence embeddings for biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 1-5). IEEE.

[17] Van-Tu, N., & Anh-Cuong, L. (2016). Improving question classification by feature extraction and selection. *Indian Journal of Science and Technology*, *9*(17), 1-8.

[18] Osadi, K., Fernando, M. G. N. A. S., & Welgama, W. (2017). Ensemble classifier based approach for classification of examination questions into Bloom's taxonomy cognitive levels. *International Journal of Computer Applications*, *162*(4), 1-6.

[19] Ray, S. K., Singh, S., & Joshi, B. P. (2010). A semantic approach for question classification using WordNet and Wikipedia. *Pattern recognition letters*, *31*(13), 1935-1943.

[20] Pota, M., Esposito, M., & De Pietro, G. (2016). A forward-selection algorithm for SVM-based question classification in cognitive systems. In *Intelligent Interactive Multimedia Systems and Services 2016* (pp. 587-598). Springer, Cham.

[21] Hao, T., Xie, W., & Xu, F. (2015). A wordnet expansion-based approach for question targets identification and classification. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (pp. 333-344). Springer, Cham

[22] Xu, S., Cheng, G., & Kong, F. (2016, November). Research on question classification for automatic question answering. In *2016 International Conference on Asian Language Processing (IALP)* (pp. 218-221). IEEE.

[23] Mohasseb, A., Bader-El-Den, M., & Cocea, M. (2018). Classification of factoid questions intent using grammatical features. *ICT Express*, *4*(4), 239-242.

[24] Liu, Y., Yi, X., Chen, R., Zhai, Z., & Gu, J. (2018). Feature extraction based on information gain and sequential pattern for English question classification. *IET Software*, *12*(6), 520-526.

[25] Lu, Z., Lin, Y. R., Zhang, Q., & Chen, M. (2016, July). Classifying questions into fine-grained categories using topic enriching. In *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)* (pp. 166-174). IEEE.

[26] Liu, Y., Ju, S., Wang, J., & Su, C. (2020). A new feature selection method for text classification based on independent feature space search. *Mathematical Problems in Engineering*, *2020*.

[27] J. Wei, C. Huang, S. Vosoughi, and J. Wei, "What are people asking about COVID-19? a question classification dataset," in Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, (Online), Association for Computational Linguistics, July 2020.

[28] R. Srihari, "A hybrid approach for named entity and sub-type tagging," in Sixth Applied Natural Language Processing Conference, (Seattle, Washington,USA), pp. 247–254, Association for Computational Linguistics, Apr. 2000.

[29] S.-J. Yen, Y.-C. Wu, J.-C. Yang, Y.-S. Lee, C.-J. Lee, and J.-J. Liu, "A support vector machine-based context-ranking model for question answering," Information Sciences, vol. 224, pp. 77 – 87, 2013.

[30] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*

[31] E. Zhang, N. Gupta, R. Nogueira, K. Cho, and J. Lin, "Rapidly deploying a neural search engine for the COVID-19 Open Research Dataset," in Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, (Online), Association for Computational Linguistics, July 2020.

[32] Magge, A., Pimpalkhute, V., Rallapalli, D., Siguenza, D., & Gonzalez, G. (2020, November). UPennHLP at WNUT-2020 Task 2: Transformer models for classification of COVID19 posts on Twitter. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)* (pp. 378-382).

[33] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu, "Performance issues and error analysis in an open-domain question answering system," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, (Philadelphia, Pennsylvania, USA), pp. 33–40, Association for Computational Linguistics, July 2002.

[34] Cortes, E., Woloszyn, V., Binder, A., Himmelsbach, T., Barone, D., & Möller, S. (2020, May). An Empirical Comparison of Question Classification Methods for Question Answering Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5408-5416).

[35] Mohammed, M., & Omar, N. (2020). Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PloS one*, *15*(3), e0230442.

[36] Schneider, K. M. (2005, October). Weighted average pointwise mutual information for feature selection in text categorization. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 252-263). Springer, Berlin, Heidelberg.

[37] Alsmadi, I., & Hoon, G. K. (2019). Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications*, *31*(8), 3819-3831.

[38] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273-292.

[39] Yen, S. J., Wu, Y. C., Yang, J. C., Lee, Y. S., Lee, C. J., & Liu, J. J. (2013). A support vector machine-based context-ranking model for question answering. *Information Sciences*, *224*, 77-87.

[40] Wang, P., Xu, B., Xu, J., Tian, G., Liu, C. L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, *174*, 806-814.

[41] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[42] S. Xu, G. Cheng, and F. Kong, "Research on question classification for automatic question answering," in 2016 International Conference on Asian Language Processing (IALP), pp. 218–221, 2016.

[43] V. Krishnan, S. Das, and S. Chakrabarti, "Enhanced answer type inference from questions using sequential models," in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, (Vancouver, British Columbia, Canada), pp. 315–322, Association for Computational Linguistics, Oct. 2005.

[44] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A.Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "Cord-19: The covid-19 open research dataset," 2020.

[45] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu, "Performance issues and error analysis in an open-domain question answering system," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, (Philadelphia, Pennsylvania, USA), pp. 33–40, Association for Computational Linguistics, July 2002.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.

[47] Zhang, X. F., Sun, H., Yue, X., Jesrani, E., Lin, S., & Sun, H. (2020). COUGH: A Challenge Dataset and Models for COVID-19 FAQ Retrieval. *arXiv preprint arXiv:2010.12800*.

[48] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

[49] Romeo, S., Da San Martino, G., Barrón-Cedeno, A., Moschitti, A., Belinkov, Y., Hsu, W. N., ... & Glass, J. (2016, December). Neural attention for learning to rank questions in community question answering. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Pa pers* (pp. 1734-1745).

[50] Abdul-Mageed, M., Elmadany, A., Pabbi, D., Verma, K., & Lin, R. (2020). Mega-cov: A billion-scale dataset of 100+ languages for covid-19. *arXiv e-prints*, arXiv-2005.

[51] Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring emotions in the covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*

[52] Zarei, K., Farahbakhsh, R., Crespi, N., & Tyson, G. (2020). A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*.

[53] Sarker, A., Lakamana, S., Hogg-Bremer, W., Xie, A., Al-Garadi, M. A., & Yang, Y. C. (2020). Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. *Journal of the American Medical Informatics Association*, *27*(8), 1310-1315

[54] Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, *6*(2), e19273

[55] X. Li and D. Roth, "Learning question classifiers," in COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[56] Britton, B. K., & Graesser, A. C. (Eds.). (1996). *Models of understanding text*. Psychology Press.

[57] Blum AL, Langley P (1997) A Selection of relevant features and examples in machine learning. Artif Intell 97(1):245–271

[58] Hsu HH, Hsieh CW, Lu MD (2011) Hybrid feature selection by combining filters and wrappers.Expert Syst Appl 38(7):8144–8150