National College of Ireland

MASTERS IN ARTIFICIAL INTELLIGENCE

PROGRAMMING FOR ARTIFICIAL INTELLIGENCE

Naresh Kumar S – 23248441
Aswin Kumar G R – 23245778

Table of Contents

Topic Modeling and Semantic Classification on Queries*

1st Naresh Kumar Satish
Msc in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23248441@student.ncirl.ie

2nd Aswin Kumar G R
Msc in Artificial Intelligence
National College of Ireland
Dublin, Ireland
x23245778@student.ncirl.ie

***Abstract*—Websites like Reddit, Quora, Answers are considered to be some of the online platforms for queries to be posted and to get them answered. The proposed work suggests an approach of topic modeling over the queries using the unsupervised machine learning algorithms like Latent Dirichlet Allocation, K-means clustering methods in catergorizing these queries to their respective domains. In addition to this the study also focuses on the detection of duplicate questions that are being posted and finds an approach in classifying these questions using machine learning algorithms like Random forest classifier and XG Boost. The applications of Supervised, Unsupervised and National Language Processing are being utilized in an effective manner in achieving the topic modeling and duplicate question detection in some of the online query posting platforms.**

***Index Terms*—Latent Dirichlet Allocation (LDA), Topic Modeling, Duplicate question detections.**

I. INTRODUCTION

In this modern era, there is a huge amount of data being collected every day, structured, unstructured, semi-structured, metadata and biometric data. There are many challenges faced in handling, storing, maintaining, categorizing, securing, summarizing, extracting valuable insights by analyzing and much more. Regarding this, there is a need for computational, storage, skilled technicians, security. The proposed work aims to solve one of the critical tasks like handling unstructured data by acquiring the applications and use cases of supervised, unsupervised machine learning algorithms and natural language processing (NLP). Topic modeling is a process of discovering the hidden thematic contexts underlying in a set of documents, one of the widely known and used applications of NLP in understanding and summarizing the collections of textual data. Considering some of the online platforms like Quora, Reddit, Answers, there is lot of knowledge and information that is being gained from these platforms. And also these are the kind of platforms where the user posts his doubt as a question, irrespective of the field or domain, and gets his questions answered. The study aims at two important challenges that these platforms might be undergoing; the questions that are being posted is done by various kinds of people from various parts of the world, with the application of NLP the proposed work aims in categorizing the questions to their respective domains they belong to. In addition to this, classifying the questions as duplicate or non-duplicates by extracting the semantic meaning of the question. With this categorization and classification, it becomes more organized way of using the websites for the users and also helps in redundancy of the data. The unsupervised machine learning algorithms like K-means and Latent Dirichlet Allocation (LDA) were employed for topic modeling on the questions, X G Boost and Random Forest classification algorithms were used for the purpose of classification of the questions into duplicate or non. The datasets used for solving these two critical tasks were questions in which the dataset used for topic modeling was web scraped from the answers.com website and the dataset used for classification task was the dataset of question pairs which is a readily available dataset that was imported from the kaggle website. Some of the NLP applications were employed in tokenization, stop word removal, stemming, lemmatization and to extract relevant features from the text input, a combination of advanced natural language processing techniques and traditional feature engineering methods is employed. These attributes capture fuzzy matching scores, lexical and syntactic similarities, and other linguistic characteristics that can indicate redundant questions. And for further visualizations of the word frequency, topic distribution applications like word cloud were used. Furthermore, dimensionality reduction techniques like t-SNE are used to obtain feature space visualizations and insights into data point distribution. The project's findings have relevance for numerous applications, such as information retrieval systems, question-answering platforms, and online discussion boards. The study proposes an efficient approach in handling unstructured data and an initiation to provide a simplified user interacting approach towards these public websites. Based on the objectives of the study the questions that were formulated related to our research:

- To find effectiveness of unsupervised machine learning algorithms like clustering and LDA on topic modeling.
- Improving the duplicate question detection with the use of supervised learning

II. RELATED WORK

Identifying duplicate questions is a valuable function in many online platforms and information retrieval systems, and it has garnered significant attention in natural language processing (NLP). Numerous studies have proposed innovative techniques and approaches to address this challenge and enhance search precision and user satisfaction. The research conducted by [1] is considered seminal work in the field of duplicate question identification. They proposed a framework that utilized textual similarity features and machine learning approaches to detect similar questions on question-answering platforms and online community forums. Their methodology demonstrated promising outcomes in identifying questions addressing comparable topics. [2] explored the use of cosine similarity and TF-IDF weighting as information retrieval techniques to identify duplicate questions. The researchers found that these strategies were particularly effective in improving the accuracy of detecting duplicate questions when working with large-scale data sets. In recent years, deep learning models have proven to be highly beneficial for tasks involving natural language understanding, such as question duplication detection. [3] introduced a Siamese convolutional neural network (CNN) architecture with the aim of identifying duplicates and capturing the semantic representations of queries. Their model outperformed traditional techniques, particularly in its ability to capture nuanced semantic similarities between questions. Advancements in word embedding methods have enabled the creation of more sophisticated algorithms for detecting duplicate questions. The researchers [4] proposed a Siamese recurrent neural network (RNN) with an attention mechanism that utilizes pre-trained word embeddings to effectively detect semantic connections between queries. Their method achieved state-of-the-art performance on common benchmark datasets, demonstrating the effectiveness of deep learning approaches in this task. In addition to neural network-based techniques, ensemble learning methods, such as combining decision trees and support vector machines (SVM), have also been explored to distinguish duplicate questions. [5] developed a hybrid model aiming to enhance the robustness and generalization of systems designed to identify duplicate questions. Furthermore, the effectiveness of algorithms used to identify duplicate questions can be improved through feature engineering. The study supervises by [6] investigating the ability of various lexical, syntactic, and semantic features extracted from question pairs to distinguish between duplicate and non-duplicate questions. Their findings emphasized the importance of feature representation and selection in achieving high accuracy in the identification of duplicate questions. The literature review highlights the diverse range of approaches employed in the identification of duplicate questions, spanning from traditional machine learning algorithms to advanced deep learning architectures. Researchers have made significant advancements in enhancing the effectiveness and efficiency of duplicate question detection systems by leveraging these techniques and implementing innovative feature engineering methodologies. [7] The proposed works on clustering and topic modeling in online platforms like TWitter and Reddit. They have used LDA to do topic modeling and have included four clustering methods, and also emphasized on the importance of term frequency and inverse document frequency.It was found that clustering methods along with the neural embedding of the features showed good results. The limitations of this paper is that the OSN data was considered as a noisy data and it was challenging in comparing the results with the number of algorithms that was used. [8] LDA was used for topic modeling, the paper emphasizes on the applications of LDA in finding the semantic relation within the documents. LDA has now developed to many forms like Hierarchical LDA, Dynamic topic model. Limitations encountered in this paper was, the number of topics assigned to the LDA model must be pre defined and its inability to capture relations on words over time. [9] The study focuses on topic modeling using k means and LDA where the news article was preprocessed using the BERT model, there were many acquired like Word2Vec, Doc2Vec, and BERT models. Autoencoders were employed for dimensionality reduction. The News article were categorized using LDA and the Doc2Vec model showed good results. The only limitation was the computational cost which was very high. [10] Tf-Idf methods were employed for the text representations and for the purpose of selection of features information gain algorithm was used. K-means clustering was used for the clustering over the texts. The paper concludes that K-means is well and suitable method for topic modeling in a large corpus of texts.[11] the customer text reviews were used for topic modeling and it was done using LDA algorithm, and after labeling of the topics XG boost algorithm was used to train on the labelled data and achieved accuracy around 0.87. Sentimental analysis was also done using Random Forest Classifier and the accuracy was noticed to be 0.93. The proposed work had put forth some of the limitations, it was challenging to identify the topics from the distribution of words.

From the related work, it is noticed that LDA and clustering methods were some of the commonly used methods in topic modeling, with some of the limitations like computational cost, noisy data and pre defining the number of topics for topic modeling.

III. METHODOLOGY

In online question and answer platforms, topic modeling might provide an alternate to the users who look to answer the questions in a specific domain and likewise while considering the aspects like data storage it becomes crucial in making the data redundant hence one of the way for making the data redundant is by detecting the question pairs. Topic modeling and duplicate question detection both of these tasks are handled with the similar data. This study aims in providing an approach to the website by doing topic modeling and detecting the duplicate question in order to make the website more user interactive and providing redundancy in the data.

A. DATASET DESCRIPTION

The dataset that was used to work on topic modeling was web scraped from an online platform answers.com. This is a website where there is lot questions that are being posted and are being answered irrespective of the domain. The data was collected with the help BeautifulSoup library package, which helps in collecting the data by parsing the HTML files. There were 5000 questions which was collected through this web scraping process. And the data that was used for detecting the duplicate questions was a readily available dataset from kaggle which consisted of question pairs and labels. The dataset consisted 400,000 question pairs with labels. The web scraped data and the question pair dataset was stored in mysql database. And for the further preprocessing and analysis, the data was being accessed from the mysql database. The collection of the data was solely based on all the ethical guidelines and it was the public data that was being utilized for the proposed work

B. DATA PREPROCESSING

Data preprocessing is the initial step in our approach, which cleans and prepares the question for analysis.On analysis there were many one worded words that were collected during the web scraping process, these questions are termed to be some of the outliers present in the data. These outliers were removed from the data. From fig.1 it is evident, the presence of one word sentence which are not questions. After the removal of these outliers there were 4312 rows of questions in the database. This data was being preprocessed using the NLP packages like Natural language toolkit (NLTK) and Genism. The basic preprocessing for the web scraped data and question pairs included cleaning of the text like removing the punctuation, converting the texts to lower case. This helps in standardization of the data and also reduces the complications over the unstructured data. Tokenization of the text data was done, where the text data was split into individual words and assigned tokens. Generally the questions possess a lot of stop words and most of the words express a verb in it, hence the stop words were removed and lemmatization was done on the data. This helps in focusing on the semantic meaning of the questions. Further the words after tokenizing, removal of stop words, lemmatization, they were subjected to align in a way representing the text data in a numerical matrix form, where each column represents a unique vocabulary term and each row represents a pair of questions. This numerical representation of the queries enables the use of machine learning techniques. Using the Gensim library a dictionary was created with the preprocessed data, the words from the preprocessed data are mapped to a unique id in the dictionary, then the words that were less occuring in the questions and the words that were occuring frequent times in the questions were removed. This was done to obtain a dictionary of words which could provide a meaningful insights. Each question was then transformed to the bag of words where each tuple represents a word with its unique id and frequency.

C. FEATURE ENGINEERING

Crafting an effective duplicate question detection system heavily relies on feature engineering. To capture the diverse aspects of similarity between question pairs, we extract a variety of attributes from them. These features encompass length-based data, such as the absolute difference in length and average token length, as well as token-based features like the count of common words, common stopwords, and common tokens. Additionally, to account for approximate string matching, we collect fuzzy features including the fuzzy ratio, fuzzy partial ratio, token sort ratio, and token set ratio.

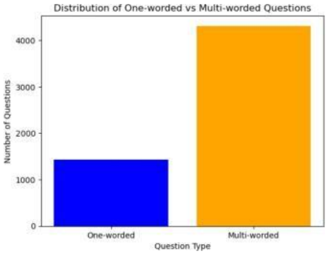


Fig. 1. Outlier analysis

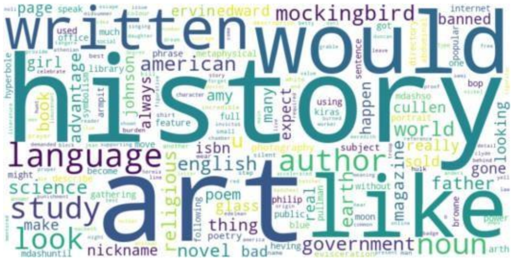


Fig. 2. word frequency analysis

D. EXPLORATORY DATA ANALYSIS

To inspect on the data that is acquired, data visualization serves as an important step in interpreting the data patterns and insights that lie inside the data. In the case of topic modeling, the term frequency of the word was analysed to determine what are the most common words that were used in the questions, this was done after the data was preprocessed. The most common words was represented visually with the help of word cloud, from the size of the word that is visualized is directly proportional to the frequency of the word in the dataset. To visualise the length of each question after preprocessing it, the length of each question was plotted using a histogram. From fig.3 it is picturizes the length of each question after preprocessing. After performing feature engineering to the detection of duplicate questions, there were series of analysis that was don on the data:

- Pairplot for Common Words Count (cwc min), Common Stop Words Count (csc min), and Token Count (ctc min) - Fig.4 pairplot visualizes the relationship between the minimum values of the common word count, common stop word count, and token count in question pairs. The color of the plot indicates whether the questions are duplicates or not.
- Pairplot for Common Words Count (cwc max), Common Stop Words Count (csc max), and Token Count (ctc max)

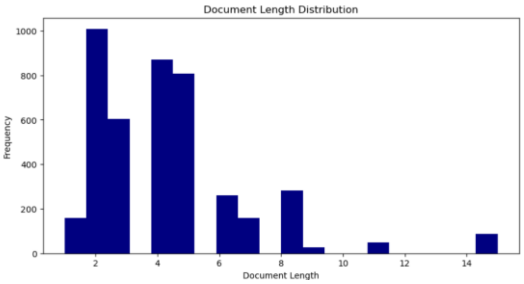


Fig. 3. question length

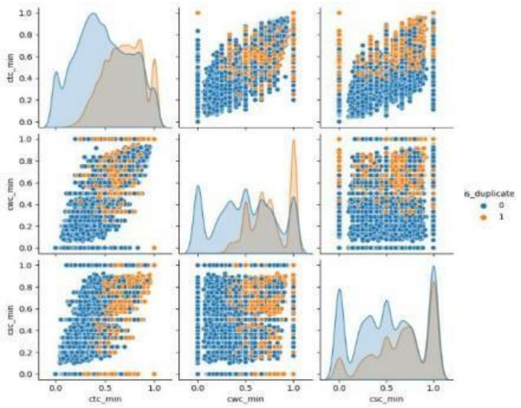


Fig. 4. pair plot for (cwc, csc, ctc)min

- - Fig.5 provided plot is similar to the previous pairplot, depicting the correlations between the highest values of the counts of common words, common stop words, and tokens in the question pairs. The color of each plot signifies whether the questions are duplicates or not.

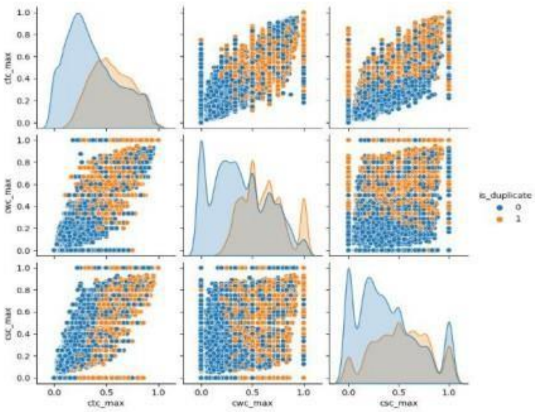


Fig. 5. pair plot for (cwc, csc, ctc)max

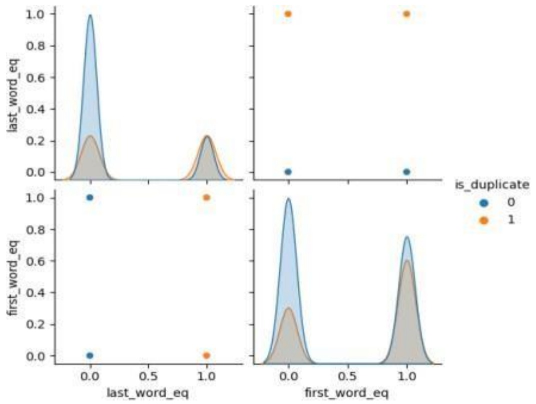


Fig. 6. pair plot for last word equality

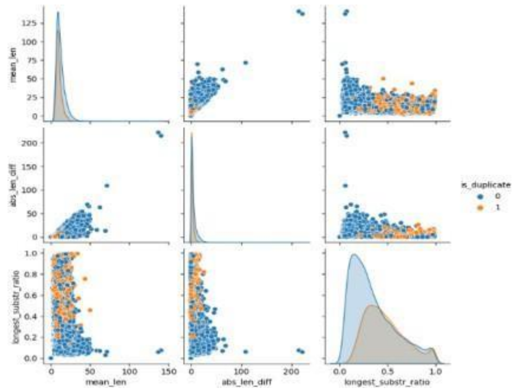


Fig. 7. pair plot for length difference

- Fig.6 Pairplot for Last Word Equality and First Word Equality - The color of the pairplot represents whether the questions are duplicates or not, and it demonstrates the relationship between the first word ”equality” and the last word ”equality” features in the question pairs.
- Fig.7 Pairplot for Mean Length, Absolute Length Difference, and Longest Substring Ratio - The pairplot displays the longest substring ratio, average length, and absolute length difference characteristics in the question pairs. The color coding indicates whether the questions are duplicates or not.
- Fig.8 Pairplot for Fuzzy Features (fuzz ratio, fuzz partial ratio,token sort ratio,token set ratio) - This pairplot shows the associations between different fuzzy matching features in question pairs, including fuzz ratio, fuzz partial ratio, token sort ratio, and token set ratio. The hue of the questions indicates whether or not they are duplicates.
- 2D TSNE Plot - It describes the use of t-distributed Stochastic Neighbour Embedding (t-SNE) to reduce the dimensionality of the features generated after data cleaning. This technique projects the features onto a two-dimensional space, and the points are colored based on whether the corresponding questions are duplicates or not.
- 3D TSNE Plot - This plot employs t-SNE for dimensionality reduction, similar to the 2D TSNE plot, but it

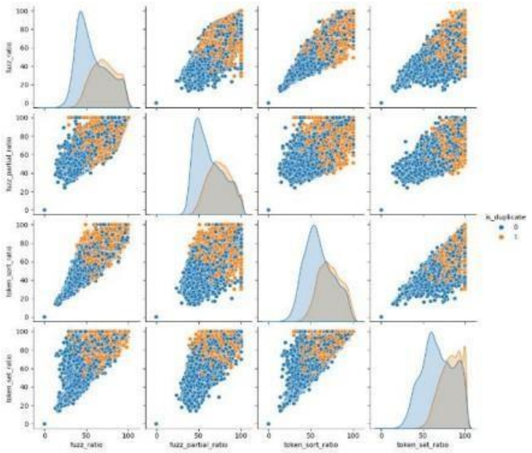


Fig. 8. pair plot for fuzzy features

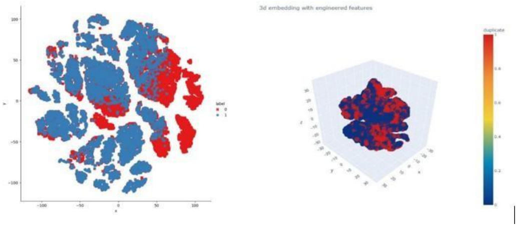


Fig. 9. 2D and 3D visualization of question pairs

presents the features in three dimensions. The points are colored based on whether the questions are duplicates or not. Fig.9 represents the 2D and 3D plot.

E. MODEL TRAINING AND EVALUATION

To categorize the questions to their domain the study approaches the use of unsupervised machine learning algorithm like LDA and K-Means,

- Latent Dirichlet Allocation (LDA): The algorithm makes use of the concept of probability, where for each question LDA assigns a probability distribution on the topics. The topics are defined to 3 during the training, in each topic the probability distribution is done over the words, the distributions are learnt by continuous iteration of the model over the data. Firstly there is creation of dictionary done from the preprocessed data, where the words are assigned unique id. After filtering of the most frequently occurring word and representing the words using Bag of words, the model is trained, after many iterations it was found that since the data is quite small, the number of topics that were assigned to LDA was 3. After the training of the model we get the topics printed where the tuples are printed with the words assigned to that topic and their probability value also printed. The model first trained by just considering the preprocessed questions as it is and got the topics probablistic output like this:
  - Topic 0: 0.037*”art” + 0.035*”written” + 0.020*”author” + 0.020*”like” + 0.019*”language” + 0.018*”english” + 0.018*”burned” + 0.017*”literature” + 0.017*”story” + 0.015*”book”
  - Topic 1: 0.038*”history” + 0.037*”u” + 0.026*”noun” + 0.017*”advantage” + 0.017*”want” + 0.017*”soundtrack” + 0.017*”movie” + 0.017*”loved” + 0.017*”flying” + 0.017*”disadvantage”
  - Topic 2: 0.040*”study” + 0.029*”american” + 0.029*”government” + 0.029*”world” + 0.028*”history” + 0.028*”science” + 0.028*”earth” + 0.028*”religious” + 0.020*”mockingbird” + 0.019*”character”

  To evaluate on the model, whether the model’s probability is done a good justice to the questions, we use the perplexity score. Perplexity score is commonly used metric to measure the model’s performance on its probability. The perplexity score the proposed work had obtained during the model training is -5.113236437050813. However the score is just a metric, in this case it was a good choice to check the assigned words to the topics manually. On manual inspection, the web scraped questions did not contain a diverse of domains instead they contained questions only about the domains related to Literature, Novels, History, English. Additionally, the LDA model was also trained on the preprocessed words by taking bigram and trigram as the word feature. And it was noticed that there was slight difference in the perplexity score, Perplexity: -5.17408295496526. However this was just a slight difference in the score it was concluded that the model could perform better in classifying the questions by considering the preprocessed questions as bigram and trigram. Below are the words assigned to the topics by considering bigrams and trigrams:
  - (0, ’0.034*”history” + 0.033*”american” + 0.033*”government” + 0.033*”u” + 0.031*”science” + 0.031*”earth” + 0.022*”character” + 0.017*”father” + 0.017*”novel” + 0.016*”future”’)
  - (1, ’0.034*”written” + 0.033*”author” + 0.024*”world” + 0.023*”would” + 0.023*”burned” + 0.023*”history” + 0.019*”thing” + 0.016*”noun” + 0.016*”like” + 0.011*”hot”’)
  - (2, ’0.047*”art” + 0.035*”study” + 0.028*”language” + 0.023*”english” + 0.023*”religious” + 0.019*”magazine” + 0.016*”mockingbird” + 0.015*”answer” + 0.012*”story” + 0.012*”little”’)

  Fig. 10 represents the word cloud the the word distribution on the topics. Perplexity score is just a metric on calculating the model performance over probability, from manual screening of the topics we can guess on the topics. Using the ’pyLDAvis’ the topics were visualized and it was noticed that the topics were well separated, the words dont overlap with other topics.
- K means algorithm was trained on the same topic mod-

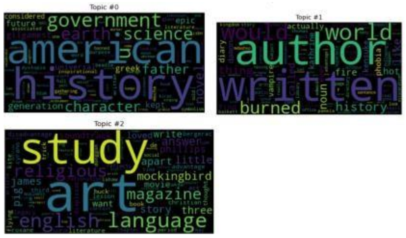


Fig. 10. word cloud for categorized topics

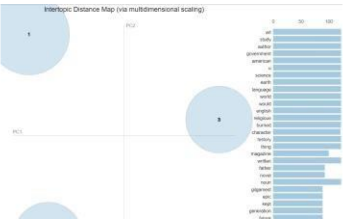
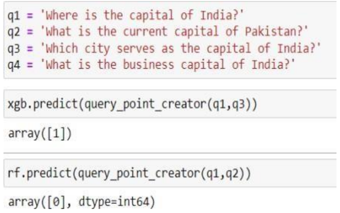


Fig. 11. topics visualized

modeling of the questions, the number of clusters assigned to model was also 3, after successful iterations on the data, the model was evaluated using the Silhouette Score which was noticed to be a 0.10391412713665818, The reason for this score might be the words which are repeated frequently and not much diversfied domains within the questions.

For detecting the duplicate questions within the question pair dataset, supervised algorithms were employed Random Forest Classifier and an XGBoost Classifier, were trained to distinguish between duplicate and non-duplicate question pairs. The models were trained using the specified features. The dataset was divided into training and testing sets in an 80-20 ratio. The models were then trained on the training set and their performance was evaluated using accuracy as the metric.On the test dataset, the RandomForestClassifier achieved an accuracy of approximately 0.78, while the XGBClassifier performed slightly better with an accuracy of around 0.79. Further analysis using confusion matrices provided more insights into the models' performance. The RandomForest Classifier accurately identified 3289 non-duplicate pairs and 1445 duplicate pairs, despite misclassifying 539 non-duplicate pairings as duplicates and 727 duplicate pairs as non-duplicates. In comparison, the XGBClassifier correctly detected 3225 nonduplicate pairs and 1519 duplicate pairs, but incorrectly categorized 603 non-duplicate pairs and 653 duplicate pairs.

- Query Point Prediction - We have developed a function that employs the same approach as the training phase to extract features and convert text inquiries into numerical representations, thereby creating a query point. Leveraging the trained classifiers, this function allows us to predict whether a new pair of questions is a duplicate. This feature can then be utilized in a practical application to automatically detect duplicate questions on a question-and-answer platform.



```
q1 = 'Where is the capital of India?'
q2 = 'What is the current capital of Pakistan?'
q3 = 'Which city serves as the capital of India?'
q4 = 'What is the business capital of India?'
```

```
xgb.predict(query_point_creator(q1,q3))
```

```
array([1])
```

```
rf.predict(query_point_creator(q1,q2))
```

```
array([0], dtype=int64)
```

Fig. 12. query point prediction

IV. CONCLUSION

The study has given a overview of the approach using of unsupervised and supervised machine learning algorithms to solve two of the critical tasks, and from gained model evaluations it is evident that model's provide good classification on the data. Regarding topic modeling the study has put forth an use case of unsupervised machine learning algorithm, for the questions that were were web scraped from the 'Answers' website. The application of NLP in preprocessing the data and visualizing the data helped in the major gain of insights. Both K-means and LDA have shown a effective way of categorizing the questions, the limitations of the use of unsupervised algorithms still exists. Some of them may include the hyperparameter tuning - which has to be changes based on the dataset. And model evaluation metrics which cannot be considered as the convenient or useful metric, manual inspection is always required on the topics. Secondly, The machine learning models utilized in classifying the duplicate questions exhibit potential in identifying whether question pairs are duplicates. We were successful in capturing the semantic resemblance between question pairs by constructing a diverse array of features, including token-based, length-based, and fuzzy features. Additionally, we incorporated Bag-of Words representations for both questions. Significant accuracy levels were achieved by the Random Forest and XGBoost classifiers, showcasing their effectiveness in distinguishing between duplicate and non-duplicate items.

V. FUTURE WORK

Future studies may continue enhancing the models' resilience and forecasting capabilities. Exploring more advanced natural language processing approaches, such as transformer-based architectures like BERT or deep neural network models like Siamese networks, which excel at identifying intricate semantic relationships in text, represents one potential direction for improvement. The models' understanding of the subtle complexities of language could be improved by incorporating contextual embeddings or pre-trained language models. Additionally, leveraging data from other sources, such as user interactions or question metadata, may provide valuable insights for enhancing model performance. The ongoing refinement and optimization of feature engineering techniques and model hyperparameters may eventually lead to even more accurate and reliable forecasts. Overall, these areas for further investigation have the potential to enhance the effectiveness of systems for detecting duplicate questions and push the limits of what is currently achievable in natural language understanding tasks.

REFERENCES

[1] Bian, J., Liu, Y., Agichtein, E. (2008). Finding the right facts in the crowd: Factoid question answering over social media. In Proceedings of the 17th international conference on World Wide Web (pp. 467-476)..

[2] Jeon, J., Croft, W. B., Lee, J. H. (2005). Finding similar questions in large question and answer archives. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 84-90).

[3] Wang, Y., Huang, M., Zhu, X., Zhao, L., Zhang, L. (2017). Bilateral multiperspective matching for natural language sentences. In Proceedings of the 26th international conference on World Wide Web (pp. 343-352)

[4] Tian, Z., He, B., Chen, Z. (2018). A deep learning approach for question matching in online forums. Information Sciences, 423, 74-84

[5] Huang, Z., Xu, W., Yu, K. (2013). Learning to rank for question-answering systems. In Proceedings of the 22nd ACM international conference on Information Knowledge Management (pp. 2383-2388).

[6] Hasan, K. S., Ahmad, A., Farooq, U., Baig, A. R. (2016). Duplicate question detection in stack overflow using supervised learning. In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 815- 820)

[7] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, "An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit," Information Processing Management, p. 102034, Apr. 2019, doi: https://doi.org/10.1016/j.ipm.2019.04.002.

[8] A. Goyal and I. Kashyap, "Latent Dirichlet Allocation - An approach for topic discovery," IEEE Xplore, May 01, 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp = arnumber=9850912 (accessed Apr. 17, 2023).

[9] Kashi Sethia, M. Saxena, M. Goyal, and R. K. Yadav, "Frame- work for Topic Modeling using BERT, LDA and K-Means," 2022 2nd International Conference on Advance Computing and Inno- vative Technologies in Engineering (ICACITE), Apr. 2022, doi: https://doi.org/10.1109/icacite53722.2022.9823442.

[10] D. Zhang and S. Li, "Topic detection based on K-means," Sep. 2011, doi: https://doi.org/10.1109/icecc.2011.6066301.

[11] M. F. Herqutanto, R. P. Zatari, and R. Sutoyo, "Topic Modeling Using LDA-Based and Machine Learning for Aspect Sentiment Analysis ," IEEE, 2023.