www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, Decembaer 2024, pp : 1556-1556

e-ISSN : 2583-1062
Impact Factor : 7.001

EFFICIENT DETECTION OF DUPLICATE QUESTION PAIRS USING MACHINE LEARNING AND NLP TECHNIQUES

Ms. Vaishali Bajpai[1]

[1]Sage University Indore, India.

E-mail: Vaishalibajpai95@gmail.com

ABSTRACT

The paper focuses on the development and implementation of a system for the detection of duplicate question pairs, using Machine Learning and Natural Language Processing techniques. Given the proliferation of forums and Q&A sites in the Internet Age, efficient ways to detect the same questions are crucially important for the quality and usability of such platforms. The goal of the paper is to devise a model that identifies correctly whether the semantic equivalence of the two input questions is correct. Various techniques in NLP are applied in preprocessing the text data, which includes tokenization, stemming, lemmatization, and finally vectorization using methods such as TF-IDF. Besides basic text preprocessing, some advanced features are extracted, which includes n-grams and cosine similarity, and keyword extraction. We further enrich our feature set by using the Fuzzy Wuzzy library to develop similarity ratios for question pairs. We further develop different models with Logistic Regression, Support Vector Machines, Random Forest, and Gradient Boosting. The paper performs a rather detailed comparison between all of these models to come up with the best one. These evaluation metrics will include accuracy, precision, recall, and the F1-score. Furthermore, tuning hyperparameters and cross-validation are part of the whole process for model performance optimization.

Keywords: Natural Language Processing (NLP), Machine Learning, Fuzzy Wuzzy Library, Text Preprocessing, Feature Engineering, Logistic Regression, Support Vector Machines (SVM), Random Forest, Gradient Boosting

1. INTRODUCTION

The rapid growth of online forums and Q&A platforms has led to an overwhelming amount of user-generated content, often resulting in a significant number of duplicate questions. These duplicates not only clutter the platforms but also hinder efficient information retrieval, making it challenging for users to find accurate answers quickly. As a result, there is a pressing need for systems that can automatically detect and handle duplicate questions to maintain the quality and usability of these platforms. This paper aims to address this issue by developing a system for detecting duplicate question pairs using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The primary objective is to create a robust model that can accurately determine whether two given questions are semantically equivalent, thus identifying duplicates effectively.[2]

To achieve this, we preprocess the text data using various NLP techniques, including tokenization, stemming, and lemmatization. The Fuzzy Wuzzy library is utilized to compute similarity ratios, further improving the feature set. Several machine learning algorithms are employed in this paper, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting. These models are evaluated based on performance metrics such as accuracy, precision, recall, and F1-score to determine the most effective approach for duplicate question detection. This report documents the methodology, implementation, and results of our approach to duplicate question pair detection. It also discusses the challenges faced and the solutions developed, offering a comprehensive overview of the paper's contributions to the field of NLP and machine learning.

2. MOTIVATION

The motivation for this paper stems from the need to enhance user experience and information retrieval efficiency on Q&A platforms by developing an automated system to accurately detect and manage duplicate questions, thereby reducing redundancy and improving overall platform quality. Key facets of our motivation include:

2.1 SCIENTIFIC INVESTIGATION

• **Advancing NLP Research:** In the process, this would further research in the area of Natural Language Processing through the examination and application of different techniques for text preprocessing and the computation of similarity measures. This broadens the horizon as to how semantic equivalence could be quantified and detected [1].

• **Understanding Language Semantics:** By analysis and checking of question pairs, the paper will enhance our understanding of the semantics of language and the small differences between similar questions. This contributes to the wider scientific enterprise of modeling human language understanding.

• **Data-driven insight:** The paper provides data-driven insight into the nature and frequency of duplicate questions, hence valuable information for any future linguistic or computational research.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, Decembaer 2024, pp : 1556-1556
e-ISSN : 2583-1062
Impact Factor : 7.001

2.2 TECHNOLOGICAL ADVANCEMENT

• **Information Retrieval Improvement:** The goal of the paper is to enhance the efficiency of information retrieval systems on Q&A platforms. This will be done by detecting duplicates for questions automatically and handling them so that users get quicker, more accurate responses.

• **Improved User Experience:** The system decreases redundancy and, as a result, clutter on the Q&A platform, improving the general user experience of finding relevant information and answers.

• **Scalability and Automation:** Developing automated duplicate detection systems empowers platforms to scale with lowered dependence on proportional increases in manual moderation effort, hence really applying AI and machine learning in practical applications in the real world [3].

2.3 LUDIC INNOVATION

- Gamification and User Engagement: Within Q&A sites, gamification would apply to the detection and management of duplicate questions by rewarding users when they mark a question as duplicate or merge similar questions, making it community-driven.
- Interactive Learning Tools: The developed technology can be linked to interactive learning tools that help users understand how to frame questions and promote high quality questions, thereby supporting educational goals on digital literacy.
- Creative Applications: The methodologies and technologies developed under this paper can inspire a variety of innovative applications beyond Q&A websites, such as in customer service chatbots, automated help desks, or collaborative knowledge bases. In responding to these motivations, the paper does not aim to just solve a specific problem but contributes to broader scientific, technological, and innovative goals in establishing the far-reaching impact of effective duplicate question detection.

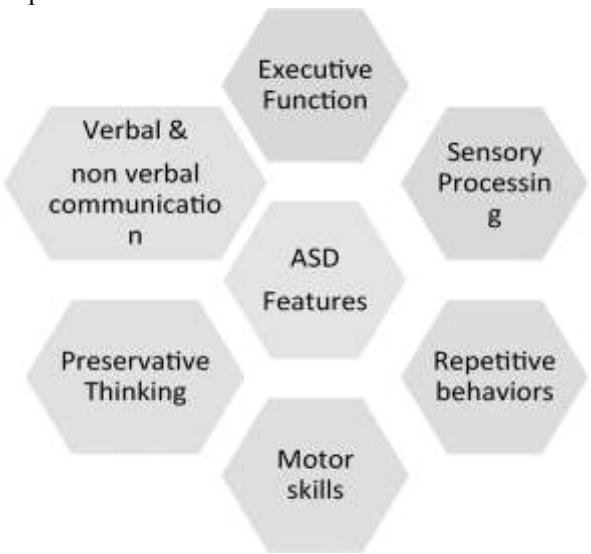

Fig.1. Features of Autism Spectrum Disorder (ASD)

3. LITERATURE REVIEW

3.1 BACKGROUND

ASD is a neurodevelopmental disorder consisting of problems in social communication, limited interests, and repetitive behaviors. Diagnosis and timely intervention are very important for a type of disorder like this. Current methods of diagnosis are mostly subjective in observation and clinical assessment, delaying identification and access to support services.

Recent technologic advances, especially those in eye movement and speech analysis, hold enormous promise for the development of innovative methods that can improve the detection of ASD. Eye tracking technologies allow inferences related to gaze patterns and visual attention, while speech analysis algorithms examine speech characteristics and prosody—both potential biomarkers for ASD.

Related Work:

- Etri Journal, "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images", Research gate, 2022[1].

This study explored using machine learning to diagnose ASD in children based on facial images. AutoML achieved the highest accuracy (around 96%) compared to other methods. This suggests AutoML is promising for ASD diagnosis, but

facial images alone might not be enough. Future work should explore combining data sources and improving model interpretability.

Mohammad Shafiul Alam, Muhammad Mahbubur Rashid, Ahmed Rimaz Faizabad,, et al. " [2]

Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum Disorder Diagnosis from Facial Images Using Explainable AI, 2023 [3]

This study creates a new method for diagnosing autism in kids using facial images and deep learning. It achieves very high accuracy (almost 99%) and uses special data tricks to make it even better. The coolest part is it explains how the model makes decisions, which helps doctors understand it.

Ying Li , Wen – Cong Huang , Pei – Hua Song , et al. " **A face image classification method of autistic children based on the two-phase transfer learning**." Front Psychol. 2023; 14: 1226470. Published online 2023 Aug 31

This study improves phone-based autism screening in kids using facial images. They make a special kind of deep learning model that works well on small phone images. This lets them get better results (over 90% accuracy) than before (around 93%). This paves the way for more accurate autism checkups on phones.

3.2 PROBLEM STATEMENT

ASDs affect millions of people worldwide, yet early detection is elusive. Early intervention is very important to improve outcome, while most cases remain undiagnosed until adulthood. Traditional diagnostic techniques are time-consuming and can be too costly, mostly based on subjective evaluation.

• The seamless integration of eye gazing and voice recognition technologies shall be developed for autism detection.

Predict autism spectrum disorder by analyzing, in real time, eye movement and speech patterns using machine learning algorithms.[4]

Evaluate the models of diagnosis for accuracy and robustness across different demographic groups and conditions.

Usability and user experience testing of the diagnosis system, with feedback taken to enhance its interaction and accessibility.

Examine the implications of early detection of autism for intervention and support for enhancing quality of life for individuals and families affected by ASD [4][5].

3.3 OBJECTIVE OF RESERCH WORK

The objectives of the research work are as follows:

- To Investigate Feasibility: Check the feasibility of integrating eye gazing and voice recognition technologies for real-time autism detection.
- To Develop Diagnosis Model: Develop machine learning models capable of predicting autism spectrum disorder based on patterns in eye movement and speech.
- To Assess Diagnosis Performance: Evaluate the accuracy, robustness, and generalization performance of diagnostic models across a wide range of demographic groups and conditions.
- Exploring User Experience with the Diagnostic System: Usability and user experience with the diagnostic system shall be explored, along with feedback for enhancement.
- Contribution to Knowledge: The research shall contribute new insights into the technology of autism detection and its implications for health and care services.
- Enable Future Applications: Build foundation for further research and future applications into early intervention and support of individuals with autism spectrum disorder.

It will improve the existing understanding of the potential and limitations of Machine Learning technology in medicine and contribute to more general scientific knowledge and technological innovation in the field of Machine Learning.[5]

4. SOLUTION APPROACH

The development of the Duplicate Question Pair Detection system involves several key stages:

- **Data Collection and Preprocessing**: Acquire a diverse dataset consisting of question pairs from Kaggle. Preprocess the dataset by removing duplicates, irrelevant information, and noise. Perform tokenization, stemming, and lemmatization to standardize the text data.
- **Feature Extraction**: Extract relevant features from the preprocessed data to represent the semantic similarity between question pairs. Utilize techniques such as n-grams, cosine similarity, and keyword extraction to capture semantic nuances and patterns. The Fuzzy Wuzzy library is used to compute similarity ratios, further enriching the feature set.

- **Model Development**: Select appropriate machine learning algorithms for the task, considering the nature of the data and the problem at hand. Develop a pipeline for model development, including data preprocessing, feature extraction, and model training. Algorithms such as Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting are employed.
- **Training and Validation**: Split the dataset into training and validation sets to assess model performance. Train the models on the training data and validate their performance using the validation set.[5]
- **Model Evaluation**: Evaluate the trained models using performance metrics such as accuracy, precision, recall, and F1-score. Conduct a comparative analysis of model performance to identify the most effective approach for duplicate question pair detection.[6]
- **Integration**: Integrate the best-performing model into the Duplicate Question Pair Detection system, ensuring seamless compatibility with existing platforms and frameworks. Develop APIs or interfaces for easy integration with other systems and applications.
- **User Testing and Iteration**: Conduct user testing to gather feedback on the system's usability, performance, and accuracy. Solicit user input to identify any issues or areas for improvement in the system's functionality and interface. Based on user feedback and performance evaluation results, iterate on the system design and implementation to address any identified issues or shortcomings. Refine model parameters, feature engineering techniques, and preprocessing methods to enhance the system's effectiveness and user satisfaction.

By following this solution approach, the aim is to develop a robust and accurate Duplicate Question Pair Detection system that improves the quality and usability of Q&A platforms by reducing redundancy and enhancing information retrieval efficiency.

4.1 PROPOSED MEHODOLOGY

- Experimental Design: Design an experimental protocol consisting of how the data will be collected, including the recruitment of participants from different demographic groups, instructions about the tasks they will perform, and how the session will be scheduled. Define variables to be measured, like eye gazing patterns, speech characteristics, and demographics of participants.
- Data Collection: Gather eye gaze and speech data in active tasks. This consists of activities created to elicit natural responses from subjects. Proper calibration of eye-tracking equipment and audio recording devices to record data.
- Design of the Interactive Task: Create standardized interactive tasks personalized to the engagement and comfort criteria for individuals going through autism assessment. Mention the conditions under which the task will be performed, stimulus presentation, and measures of performance in patterns of eye gaze and speech.
- Data Preprocessing: The collected eye gazing and speech data needs to be cleaned of noise and artifacts. Filtering techniques and noise reduction methods are applied to enhance the quality of data for further analysis.
- Feature Extraction: Features of eye gazing and speech data from the preprocessed data must be extracted that may indicate autism spectrum disorder. Compute metrics, including gaze fixation duration, speech fluency indices, and prosodic features from the segmentation of data.
- Model Development: Come up with machine learning models that support the prediction of autism spectrum disorder based on the extracted features. Experiment with various classification algorithms like Support Vector Machines, Random Forests, and Logistic Regression to arrive at the best model.
- Train and Validate Models: Train the predictive models on labeled eye gazing and speech data along with clinical diagnosis. Cross-validation techniques will be used to evaluate the performance of the models, tune hyper-parameters, and avoid overfitting.[7][8][9]
- Evaluation and Analysis: Calculating model performance metrics, such as accuracy, sensitivity, specificity, and AUC. Now, analyze the results to draw inferences regarding the relationship between eye gaze, speech patterns, and autism spectrum disorder to conclude the feasibility and efficacy of using these modalities for ASD prediction.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, Decembaer 2024, pp : 1556-1556
e-ISSN : 2583-1062
Impact Factor : 7.001

4.2 Methodology Flow Chart:

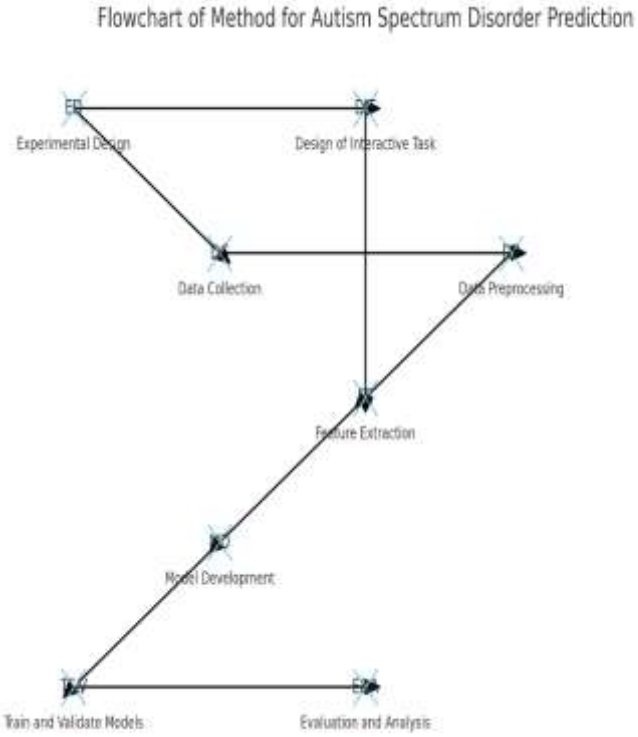

Fig.2. flowchart representing the method for autism spectrum disorder (ASD) prediction

5. RESULT

5.1 DATASET

A well-filtered and representative dataset is inert in ensuring the model's robustness and applicability. Following is a discussion of the dataset in this regard:

1. Dataset Selection Criteria and Representation

The selection criteria of datasets for the ASD detection paper include the development of a local dataset. It will be achieved by acquiring videos and images of autistic children from prominent organizations. The dataset needs to be vast and includes a wide variety of people with ASD so that it rightly represents and models can be trained robustly from it. Sources from trusted organizations like NGOs working in autism research and support will have a varied pool of behaviors and characteristics normally associated with autism. This method enhances the diversity and genuineness of the dataset, thereby helping to train a model for more accurate and efficient ASD detection using machine learning and deep learning techniques.

2. Eye Tracking for ASD Screening:

Eye tracking is a noninvasive technique that could register a person's gaze positions in real-time. Eye-tracking data present one of the potential biomarkers for very early ASD detection, as people with autism display rather unique patterns of visual attention and oculomotor function.

Fig.3. Eye Tracking for ASD Screening

3. Speech-based Analysis for ASD Detection: Speech and communication deficits have been described as very common symptoms of ASD. The speech patterns and characteristics can be analyzed to help in early detection and diagnosis of ASD.

4. Multimodal Approach:

Combining the information derived from eye-tracking and speech analysis will provide an ensemble approach to more accurate diagnostics in ASD. Such integration of multiple data sources—gaze, facial expressions, and speech—will enhance the discrimination between those with and without ASD.

6. CONCLUSION

Specifically, advancements in eye gaze and speech technologies could help enormously towards the diagnosis and support of ASD. Each technology brings objectivity into the assessment, individualized intervention, and remote monitoring, which enhances the quality of care and support for individuals with ASD and families.

Eye gaze technology would offer one the potential for early detection of ASD by analyzing atypical patterns of eye movements in social interactions. It gives a more objective and quantifiable measure of the social communication deficits of ASD, thus enabling clinicians to track progress over time by customizing interventions to address specific problems.

Complementing this is speech recognition technology that monitors language development and analyzes the social communication behaviors of people with ASD. This is according to speech recognition systems, which provide key information on the level of communication competency of a subject through conversation transcription and speech, turn-taking, and pragmatic patterns of language use, and help in developing tailored interventions.

In addition, integration with other modalities—like facial expression analysis—is another area that can further the comprehensive assessment of social communication skills in individuals with ASD. Furthermore, the remote monitoring options for these technologies enable assessments within naturalistic settings and, therefore, earlier possibilities of intervention and more frequent follow-up assessments. In general, synergizing eye gaze technology and speech recognition has some promising potential in autism diagnosis and the facilitation of ASD patients with the development of social skills in communication. With the use of such novel methods, professionals like clinicians or therapists or educators will be able to provide appropriate care to the concerned individuals, thus helping to improve their quality of life and providing them with social inclusion**.**

7. FUTURE WORK

7.1 Eye Gaze Technology:

Early Detection: Eye gaze technology can be utilized in the early detection of autism spectrum disorder (ASD). Research has shown that people with ASD may reveal atypical patterns within their gaze, such as reduced eye contact or unusual fixation on certain objects or patterns. Analyzing these patterns, eye gaze technology can aid in detecting children at risk for ASD early in their age.

7.2 Objective Assessment: The conventional diagnostic techniques of ASD are largely based on subjective observation and clinical evaluation. Eye gaze technology allows for the measure of social communication deficits associated with ASD in a more objective and quantifiable manner. Eye movement tracking and analysis during social interactions provide information about an individual's social engagement and communicative behaviors.

Personalized Interventions: Eye gaze technology can help in the development of personalized interventions for individuals diagnosed with ASD. To this day, understanding a person's unique gaze patterns and preferences helps therapists and educators devise interventions that will help zero in on very specific areas of communication challenges and development of social skills.

Longitudinal Monitoring: Eye gaze technology enables monitoring social communication skills in individuals with ASD across time. It allows one to follow changes in eye gaze patterns to establish the effectiveness of interventions and identify improvement or regression.

7.3 Speech Recognition:

Language Development Monitoring: Speech recognition technology can track language development in children with ASD. These acoustic and linguistic parameters that can be derived from speech patterns, such as vocabulary size, syntax, and pragmatic language, may show a person's language skills and pinpoint probable delays or atypicalities indicative of ASD.

7.4 Social Communication Analysis: Speech recognition technology can aid in the analysis of social communication behaviors for individuals with ASD. A speech recognition system, in transcribing and analyzing conversational

interactions, permits the identification of speech patterns, turn-taking situations, and pragmatic uses of language that may indicate typical ASD-related communication difficulties.

7.5 Multimodal Assessment: Speech recognition integrated with other modalities, such as eye gaze tracking and facial expression analysis, can be used for a more detailed assessment of social communication skills in individuals with ASD. This combination can hence provide an understanding of the person's strengths and weaknesses in communication.

It facilitates remote monitoring of social communication skills and thus allows clinicians to evaluate individuals with ASD in naturalistic settings outside a clinic. The remote monitoring capability that this speech recognition technology provides enables earlier intervention and more frequent monitoring of progress toward better outcomes for those diagnosed with ASD.

Both gaze technologies and speech recognition have some potential to contribute to autism detection and, consequently, support people with ASD in acquiring social communication competencies. These technologies hold potential for the enhancement of quality care and support through the provision of objective assessments, individually-tailored interventions, and telemonitoring.

8. REFERENCE

[1] (PDF) Efficient Deep Learning-Based Data-Centric Approach for Autism Spectrum...
[2] https://www.researchgate.net/publication/373520712_Efficient_Deep_Learning-Based_Data-Centric_Approach_for_Autism_Spectrum_Disorder_Diagnosis_from_Facial_Images_Using_Explainable_AI
[3] Mohammad Shafiul Alam • MDPI • Aug 29, 2023
[4] https://link.springer.com/article/10.1007/s44196-024-00491-yGhosh, T.; Al Banna, M.H.; Rahman, M.S.; Kaiser, M.S.; Mahmud, M.; Hosen, A.S.M.S.; Cho, G.H. Artificial Intelligence and
[5] https://ouci.dntb.gov.ua/en
[6] https://www.classace.io/answers/code-me-learning-ai-system
[7] https://www.classace.io/answers/code-me-learning-ai-system
[8] https://www.ijraset.com/research-paper/autism-detection-from-facial-images
[9] https://link.springer.com/article/10.1007/s44196-024-00491-y
[10] https://ouci.dntb.gov.ua/en/?backlinks_to=10.3390/s20236762
[11] https://ouci.dntb.gov.ua/en/works/732xqy69/