National College of Ireland

Quora Insincere Question Classification with word Embedding Algorithms

MSc Research Project
Data Analytics

Namrata Suryawanshi
Student ID:x21197741

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
MSc Project Submission Sheet
School of Computing

Student Name:	**Namrata Ashok Suryawanshi**
Student ID:	**X21197741**
Programme:	**Msc Data Analytics**
Year:	**2022-2033**
Module:	
Supervisor:	**Teerath Kumar menghwar**
Submission Due Date:	
Project Title:	**Quora Insincere Question Classification with word Embedding Algorithms**
Word Count:	**7966**
Page Count:	**31**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	**Namrata Suryawanshi**
Date:	**18/09/2023**

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Quora Insincere Question Classification with Word Embedding Algorithms

Namrata Suryawanshi

x21197741

Abstract

This report presents a comprehensive study on the application of DL models for predicting the sincerity of questions on the Quora platform. The primary objective is to differentiate between sincere and insincere questions, facilitating content moderation and enhancing user experience. The study explores various model architectures, including CNN+LSTM+GRU and BILSTM, with and without attention mechanisms and GloVe. The models are implemented and evaluated using performance metrics such as precision, accuracy, F1-score, recall, sensitivity and specificity. Among the models examined, the BILSTM model with an added layer of attention mechanism and pre-trained GloVe embeddings emerges as the best-performing model, achieving an impressive validation accuracy of 89.10%. This highlights the significance of attention mechanisms and pre-trained embeddings in enhancing model performance. The findings demonstrate the effectiveness of DL approaches for classifying question sincerity on Quora. The results hold significant implications for content moderation and user engagement on social media platforms. Additionally, the study identifies potential areas for further research, such as exploring different embeddings, ensembling techniques, and addressing class imbalances to improve model performance. This research contributes useful insights regarding the use of DL techniques for content analysis and classification on social media platforms and sets the stage for future advancements in this domain.

Keywords: CNN+LSTM+GRU, BILSTM, Attention Mechanism, Pre-trained Word Embeddings

1 Introduction

In recent times, the Internet has brought about swift transformations in people's lifestyles (Shrivas, 2021). Among the various online platforms, Quora stands out as a dynamic space that fosters mutual learning. It offers users the opportunity to ask questions and engage with a

community that provides valuable perspectives and well-founded responses (Chittari et al., 2022). Online question-and-answer platforms have become integral to the sharing of information among people across the globe. These platforms serve as valuable mediums for knowledge exchange and community engagement. However, despite their benefits, these platforms often face challenges associated with the quality of content. Many websites prioritize content generation without considering its quality, leading to issues such as online hostility, racism, and spam. The proliferation of such toxic behaviours disrupts the user experience, infringes upon platform guidelines, and negatively impacts online communities' norms and health. This phenomenon can even result in the creation of polarized groups due to the propagation of misinformation. Boasting a massive user base, Quora has become a prominent question-and-answer forum, attracting approximately 300 million active monthly users (Gandhi, 2021). This popularity underscores its role as a hub for collaborative learning and information sharing. As a response to this challenge, researchers have explored methods to mitigate biased, toxic, and insincere content on these platforms. One area of focus has been the identification and classification of insincere questions. The Quora Insincere Questions Classification research challenge exemplifies this effort. Quora, a prominent question-and-answer website, has been grappling with insincere or malicious questions that aim to deceive or provoke other users. Detecting and managing such false content is essential to maintain the platform's integrity, user trust, and quality of interactions. Several studies have contributed to addressing the issue of insincere questions. Transformer-based models, such as BERT, offer attention-based mechanisms that excel in handling noisy datasets, allowing for better generalization. This research aims to fine-tune pre-trained models using the Quora Insincere Questions Classification dataset to enhance the classification of insincere questions. The remainder of this report is organized as follows: Section 2 provides an overview of related research, Section 3 details the methodology, Section 4 presents the design specification, Section 5 details the implementation, Section 6 discusses the experimental results and evaluation and finally, Section 7 offers conclusions and suggestions for future work.

1.1Background

The digital age has witnessed an unprecedented surge in online interactions and information-sharing platforms, leading to the proliferation of user-generated content. Among these platforms, question-and-answer websites like Quora have emerged as prominent avenues for knowledge exchange. However, this growth has also brought to the forefront a pressing issue

– the presence of insincere and toxic content that not only tarnishes the user experience but also poses potential harm to online communities. Insincere questions, characterized by their intention to deceive, manipulate, or provoke, have become a significant challenge in maintaining the quality, credibility, and safety of platforms like Quora. Detecting and filtering out such content is a critical endeavor to uphold the integrity of online communities. To address this challenge, researchers and practitioners have turned to the realm of NLP and ML to develop robust models capable of differentiating between genuine and insincere questions. One notable advancement in NLP is the attention mechanism, which has revolutionized how models process and understand language. Attention mechanisms allow models to assign varying degrees of importance to different parts of input text, enabling them to capture intricate relationships and dependencies between words. This mechanism has demonstrated exceptional effectiveness in a wide range of NLP tasks, from machine translation to sentiment analysis. The BiLSTM architecture, on the other hand, has gained prominence for its ability to capture contextual information from both preceding and succeeding tokens in a sequence. This bidirectional context makes BiLSTM well-suited for tasks that require a comprehensive understanding of sentence semantics and sentiment. Combining the power of attention mechanisms with the context-capturing capability of BiLSTM presents a novel and promising approach to the challenge of insincere question detection. While attention mechanisms have been explored in various NLP domains, their application to the specific context of identifying insincere content remains relatively underexplored. By introducing attention mechanisms into the BiLSTM framework, the study aim to enhance the model's capacity to discern nuanced linguistic cues and subtle patterns indicative of insincerity. In essence, our focus on integrating attention mechanisms into the BiLSTM model emerges from the need to harness the latest advancements in NLP to tackle the persistent issue of insincere content. By building upon the collective strengths of attention mechanisms and BiLSTM, the study aspire to develop a more accurate, efficient, and insightful model for insincere question classification, contributing to the creation of safer and more constructive online interactions.

1.2Motivation

In the realm of NLP, the evolution of advanced models and techniques has revolutionized how the study approach various linguistic tasks. One such innovation is the attention mechanism, a pivotal concept that has significantly enhanced the performance of NLP models. As the study delve into the challenge of identifying insincere questions on platforms like Quora, the study

recognize the potential of leveraging attention mechanisms to elevate our classification methods. The attention mechanism introduces an element of contextual awareness by allowing the model to concentrate on certain aspects of the input text, thereby capturing intricate relationships between words and phrases. This is particularly valuable for tasks involving sentence semantics, sentiment analysis, and classification – all of which are central to discerning insincere questions. By incorporating attention mechanisms, the study aim to empower the BiLSTM model with an enhanced ability to grasp nuances and subtleties in the language, ultimately leading to more accurate and robust predictions. Our focus on integrating the attention mechanism into the BiLSTM architecture is novel in the context of insincere question detection. While attention mechanisms have been applied extensively in various NLP tasks, their application to the specific challenge of identifying insincere content is relatively unexplored. By infusing attention into the BiLSTM model, the study anticipate uncovering previously unnoticed patterns and features that can shed light on the intricacies of insincere question formulation. Moreover, the synergy between attention mechanisms and BiLSTM's bidirectional processing holds great promise. BiLSTM's ability to capture both preceding and succeeding contextual information aligns seamlessly with the attention mechanism's goal of identifying salient tokens. The fusion of these two techniques has the potential to unlock new dimensions of understanding in the complex landscape of insincere questions. In conclusion, our motivation for incorporating attention mechanisms into the BiLSTM model stems from the desire to introduce an innovative approach to the task of insincere question detection. By harnessing the power of attention mechanisms within a bidirectional context, the study seek to enhance the precision, recall, and overall efficacy of our model, ultimately contributing to the advancement of insincere content identification and fostering safer and more informative online communities.

1.3Research Questions and Objectives

1.3.1 Research Questions

RQ1: How can deep learning models be utilized to predict the sincerity of questions on the Quora platform?

RQ2: What are the differences between sincere and insincere questions, and how can these differences be effectively captured using DL techniques?

RQ3: How do different model architectures, including CNN+LSTM+GRU and BILSTM,

perform in differentiating between sincere and insincere questions?

RQ4: What is the impact of incorporating attention mechanisms and GloVe on the performance of DL models in question sincerity prediction?

1.3.2 Research Objectives

1. To apply DL models for predicting the sincerity of questions on the Quora platform.
2. To differentiate between sincere and insincere questions to enhance content moderation and user experience.
3. To explore various model architectures, including CNN+LSTM+GRU and BILSTM, with and without attention mechanisms and pre-trained word embeddings (GloVe).
4. To evaluate the models using performance metrics including precision, accuracy, F1-score, recall, specificity, and sensitivity.
5. To identify the most effective DL model architecture for classifying question sincerity on Quora.

1.4 Research Gaps

In this study, while substantial insights are gained into the application of DL models for question sincerity prediction on Quora, certain research gaps remain. Firstly, despite the focus on DL architectures, the impact of data preprocessing techniques on model performance is relatively underexplored. Secondly, the investigation primarily concentrates on performance metrics, with limited attention to the interpretability of the models' decisions. Additionally, the study largely operates within the context of the Quora platform, potentially limiting the generalizability of the findings to other social media domains. Moreover, the effects of various attention mechanisms on model interpretability and effectiveness warrant further exploration. Lastly, the research could delve deeper into the ethical implications of content moderation decisions based on DL predictions, considering potential biases and consequences. Addressing these gaps would enrich the understanding of DL's application in question sincerity classification and foster more robust, ethically sound solutions in social media content moderation.

2 Literature Review

This literature review aims to evaluate and juxtapose previous studies addressing this challenge through ML methods. Listed chronologically, starting with the latest, these works endeavor to tackle the issue systematically.

2.1 Research based on LSTM

(Aslam et al., 2021) proposes a DL-based solution for insincere question classification on question-answering websites. The approach utilizes LSTM neural networks for text classification. The dataset is collected from Quora through Kaggle, preprocessed, and used for training the models in MATLAB. Various feature engineering techniques are considered to optimize model performance, and the F1 score is used as the evaluation metric due to class imbalance. LSTM outperforms other models in terms of F1 score, offering valuable insights for question-asking forums. This research proposes a strategy for dealing with the problems of duplicate and disingenuous inquiries on Quora NLP and DL techniques. There are five distinct word embeddings. are employed for both problems, with BiLSTM and BiGRU architectures featuring attention mechanisms for question pair identification, and Siamese MaLSTM architecture for insincere question classification. The models' performance is evaluated using precision, accuracy, F1 Score and recall. For Quora Question Pairs Identification, Paraphrase-MiniLM-L6-v2 + Siamese MaLSTM achieves the highest accuracy of 90% and the highest F1 score of 0.89. For Insincere Questions Classification, FastText + BiLSTM + BiGRU achieves the highest accuracy of 95% and the highest F1 score of 0.82. Comparison with baseline models demonstrates the superior performance of the proposed approach (Gontumukkala et al., 2022). This research aims to categorise dishonest Quora questions using pre-trained state-of-the-art language models, namely BERT, XLNet, StructBERT, and DeBERTa. To overcome high computation requirements, these models are trained on three data samples. The study explores whether limited computation can achieve remarkable results in understanding sincerity. Evaluation metrics include balanced accuracy, macro F1-score, and weighted F1-score. The authors propose utilizing transformer-based models to classify sincere and insincere questions with limited computation resources. (Chakraborty et al., 2022) compares XLNet, BERT, DeBERTa and StructBERT models and evaluates their performance using balanced accuracy, macro F1-score, and weighted F1-score. The metrics used to assess model performance are balanced accuracy, which achieves 80%, macro F1-score at 0.83, and weighted F1-score at 0.96. Among the four models, DeBERTa stands out with the highest scores, demonstrating its effectiveness in classifying insincere questions on Quora. The paper introduces the "Deep

Refinement" pipeline, which utilizes state-of-the-art methods like capsule networks and attention mechanisms for obtaining information from sparse data. The pipeline is applied to divide content into two categories: genuine and fake, with a focus on community question-answering systems. The proposed method aims to enhance monitoring and information quality by effectively classifying insincere questions. The authors propose the "Deep Refinement" pipeline, incorporating capsule networks and attention mechanisms for text classification in community question-answering systems. The method aims to improve the identification of insincere content, ensuring enhanced monitoring and information quality. The proposed method ((Jain et al., 2020) is evaluated using the F1 score, achieving a high F1 score of 0.978. This research focuses on filtering insincere and spam content on Online Social Networks (OSNs) using a real-world dataset from Quora on Kaggle.com. Various mechanisms and algorithms are evaluated for insincerity prediction, including preprocessing and analysis models. The study also investigates the cognitive efforts made by users in writing their posts to enhance prediction accuracy (Ramahi et al., 2020).

2.2 Research based on CNN

The research suggests a comprehensive approach to filter insincere and spam content on OSNs using a real-world dataset from Quora. Different preprocessing and analysis models are evaluated to identify the best mechanisms for insincerity prediction. Additionally, the study explores the role of cognitive efforts in post-writing and its impact on prediction accuracy. The evaluation metrics used to assess the models' performance include insincerity prediction accuracy. The best models are reported based on their accuracy in detecting insincere and spam content on OSNs. This study suggests a multi-layered CNN model to identify and minimize insincere questions on the Quora QA platform. Two embeddings, Skipgram and Continuous Bag of Word model, along with a pre-trained GloVe embedding vector, are utilized for system development. The model requires only the question text, eliminating the need for manual feature engineering. The paper introduces a multi-layer CNN model for detecting insincere questions on Quora. Embeddings from Skipgram, Continuous Bag of Word, and pre-trained GloVe vectors are used. The model achieves high performance with an F1-score of 0.98 for the best case, surpassing previous works. The model's performance is assessed using the F1-score, which demonstrates its effectiveness in identifying and minimizing insincere questions on the Quora QA platform (Roy and PK, 2020). The paper proposes a system to predict if a question on Quora is insincere or sincere by utilizing a significant amount of data from the platform.

Different approaches are employed to develop models that take the question text as input and output a binary value (0 or 1) indicating its sincerity. The goal is to ensure user-friendly and safe content on Quora (Singh and Mona, 2019). The proposed system aims to classify questions on Quora as sincere or insincere. Various models and approaches are employed using a substantial amount of data from the platform. The evaluation metrics focus on precision, accuracy, F1-score and recall to determine the effectiveness of the models in predicting question sincerity. The evaluation metrics include accuracy, precision, recall, and F1-score to assess the models' performance in distinguishing sincere and insincere questions on Quora. The goal is to create a system that effectively flags and handles insincere content to maintain a safe and user-friendly environment on the website. The paper presents a ML model for classifying questions as 'sincere' or 'insincere' on QA forums. The dataset from Quora via Kaggle consists of over 1.3 million labelled examples for training and 300 thousand unlabeled examples for testing. Artificial recurrent neural network architectures like LSTM and GRU are implemented. Cross-validation is done using 10% of the training data, achieving an F1 score of 0.6913 with a threshold of 0.35. The paper proposes a machine-learning model for question classification on QA forums using LSTM and GRU architectures. The dataset from Quora is utilized for training, and cross-validation is performed. The evaluation metric used is the F1 score, achieving a value of 0.6913 with a threshold of 0.35 for distinguishing 'sincere' and 'insincere' questions (Gate et al., 2019). The paper addresses the problem of Insincere Questions Classification on content-based websites like Quora, Reddit, etc., by fine-tuning four advanced models: BERT, RoBERTa, DistilBERT, and ALBERT. These models leverage transfer learning in NLP and the power of transformers to detect and prevent toxic and insincere content from proliferating online (Rachcha et al., 2019). The proposed approach involves fine-tuning four SOTA models, BERT, RoBERTa, DistilBERT, and ALBERT, for Insincere Questions Classification on content-based websites. These models leverage transfer learning and NLP innovations to address the issue of toxic and divisive content, ensuring a safer and more sincere online environment. The evaluation metrics sometimes used evaluate the performance of fine-tuned models are not explicitly mentioned in the provided text. However, typical evaluation metrics in such classification tasks may include accuracy, precision, recall, and F1-score, among others. The paper presents a DL-based system for fine-grained insincere question classification in the CIQ track of FIRE 2019. The system employs a checkpoint ensemble and combines different embeddings per-ensemble to improve performance. The evaluation metric used to assess the system's performance is the F1 score, which achieved a value of 67.32%, resulting in securing the first position in the CIQ track of FIRE 2019 (Das et al., 2019).

3 Research Methodology

The research methodology adopted for this binary classification study focuses on predicting the sincerity of questions asked on Quora. The overall approach involves analyzing the characteristics of questions to distinguish between sincere and insincere ones. To achieve this, the study utilizes the provided dataset, comprising unique question identifiers (qid), corresponding Quora question texts (question_text), and target labels (target) denoting insincere questions (target = 1) or sincere questions (target = 0). The primary objective is to develop a robust binary classification model capable of identifying insincere questions accurately. The methodology encompasses data preprocessing, where text cleansing and tokenization are performed to ensure data readiness. During this stage, special attention is paid to handling sensitive content, such as sexual content involving incest, bestiality, or paedophilia, to ensure ethical compliance and respect for users' safety. The extracted features from the question texts are then used to train and validate the predictive model. A pre-trained language model is employed to capture nuanced patterns, non-neutral tones, rhetorical implications, and disparaging language indicative of insincerity. Additionally, the model accounts for characteristics that are not fixable or measurable, as well as those based on incorrect information or irrational assumptions. The training process involves optimizing hyperparameters and addressing class imbalance to enhance the model's effectiveness. The study acknowledges the potential noise in the ground-truth labels, as they may not be perfect due to the manual nature of labelling. To mitigate this, the researchers conduct a careful analysis of misclassified instances to identify potential areas of improvement for the model. The research also recognizes the dataset's non-representativeness of the broader Quora question distribution. This limitation is addressed by emphasizing the focus on the specific characteristics that indicate insincerity while acknowledging that the dataset may not fully capture all variations of sincere and insincere questions on Quora. By implementing this research methodology, the study aims to contribute to content moderation on the Quora platform, improving the overall user experience and fostering a more respectful and engaging community. The development of an accurate and robust binary classification model for identifying insincere questions holds the potential to enhance the platform's integrity and trustworthiness, making it a valuable resource for users seeking genuine information and helpful answers. Additionally, the research methodology emphasizes ethical considerations,

ensuring responsible handling of potentially harmful content and promoting a safer and more inclusive online environment.

3.1 Methodology

This study's methodology section describes the overall methodology and techniques used to attain the research objectives of predicting the sincerity of questions asked on Quora. The study revolves around binary classification, where the goal is to differentiate between sincere and insincere questions based on certain characteristics. The study have used a single dataset in this study for binary classification. The dataset is obtained from the "Quora Insincere Questions Classification" competition on Kaggle. The goal is to develop a model that can accurately differentiate between sincere and insincere questions.

3.2 CRISP-DM Modules

The CRISP-DM technique is made up of six parts. Some modules provide two-way routes that allow you to change any prior step as needed. The six modules are listed below.:

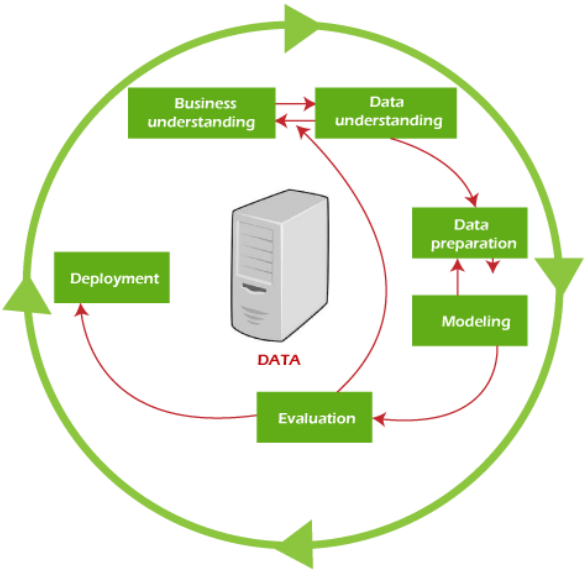


Figure 3.1: CRISP-DM Methodology

Source: "What Is CRISP in Data Mining - Javatpoint." Www.javatpoint.com, www.javatpoint.com/what-is-crisp-in-data-mining.

1. Business Understanding: The research methodology revolves around addressing the crucial issue of forecasting the truthfulness of comments on the Quora site. With the rapid proliferation of internet sites such as Quora, it becomes imperative to distinguish between sincere and insincere questions for effective content moderation and user experience enhancement.

2. Data Understanding: The primary dataset used for this study contains essential components,

including unique question identifiers (qid), the corresponding question texts (question_text), and target labels (target) denoting the sincerity status of each question. The target labels classify questions as either insincere (target = 1) or sincere (target = 0).

3. Data Preparation: The data preprocessing phase involves text cleansing and tokenization to ensure that the data is suitable for analysis. Special attention is given to handling sensitive content, reflecting ethical considerations and ensuring user safety. The extracted features from the question texts serve as inputs for model training and validation.

4. Modeling: The core of the methodology lies in developing a robust binary classification model. The approach involves utilizing a pre-trained language model to capture intricate patterns, tones, and language characteristics that indicate insincere questions. The model also accounts for characteristics that are not measurable or based on false information.

5. Evaluation: Model evaluation is a critical module that assesses the effectiveness of the developed classification model. Various performance metrics, including precision, accuracy, F1-score, recall, specificity, and sensitivity, are employed to measure the model's predictive capability.

6. Deployment: While the deployment phase is not explicitly mentioned, the research methodology's ultimate goal is to contribute to content moderation on the Quora platform. The accurate and robust binary classification model, once developed, can be integrated into the platform to identify insincere questions and enhance user engagement.

3.3 Data Visualization

The data visualization section of this study presents visual summaries and insights derived from the collected dataset. A pie chart was used to showcase the distribution of the target column, indicating that 50% of the questions were labelled as sincere and the other 50% as insincere. Another visualization employed WordClouds to display the most frequent words for each target class. The WordCloud for insincere questions highlighted words like "women," "black," "Trump," "black-white," and "Indian," potentially indicative of inflammatory or divisive content. In contrast, the WordCloud for sincere questions emphasized words like "best," "will," and "all," reflecting the pursuit of genuine and helpful answers. Additionally, a histogram was utilized to analyze the distribution of sentence lengths in the dataset, providing insights into the variability and frequency of sentence lengths. This data visualizations aid in understanding the characteristics of the dataset, guiding subsequent analysis and model development for the

binary classification task of predicting question sincerity on Quora.

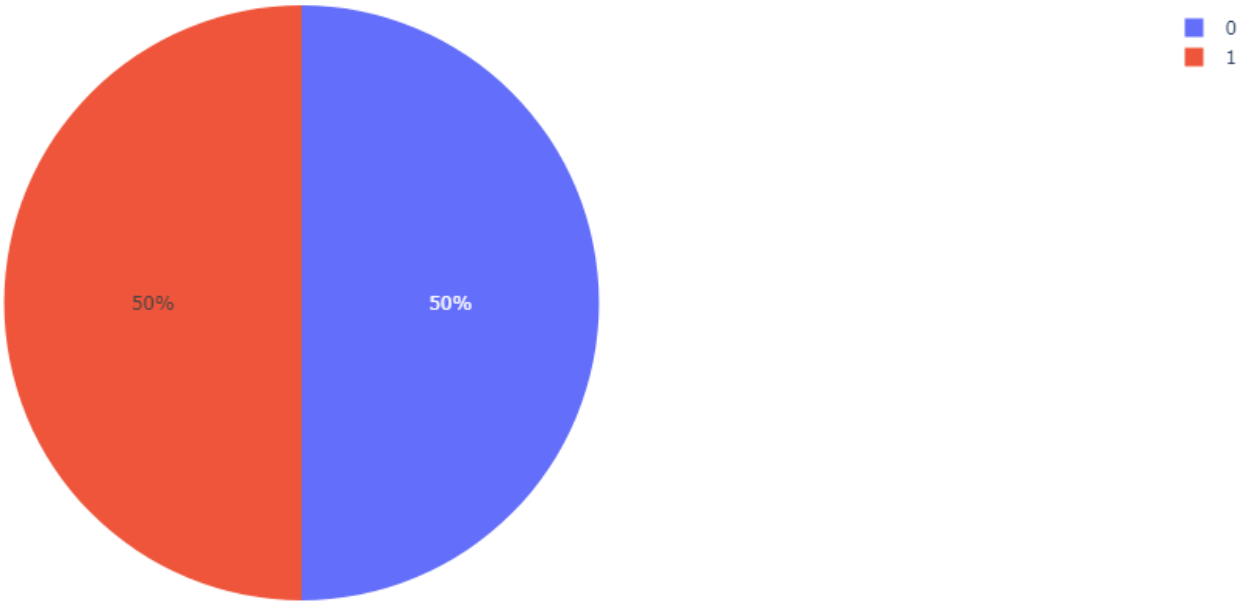
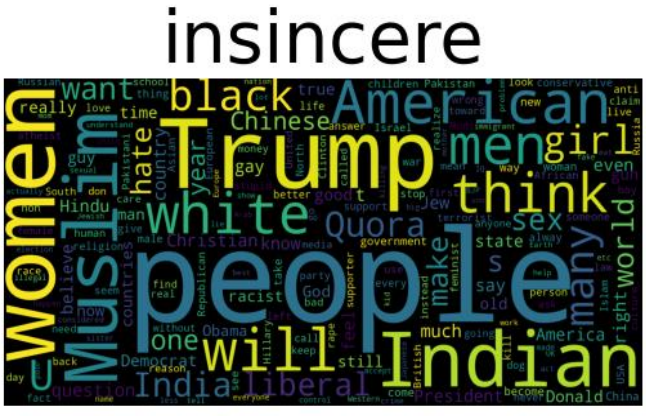


Figure 3.2 Pie Chart- Count Plot of Target Column

The pie chart in Figure 3.2 displays the distribution of the target column in the dataset. The target column represents whether a question asked on Quora is sincere or insincere. In this binary classification task, a value of 0 denotes a sincere question, while a value of 1 represents an insincere question. The pie chart is divided into two segments, each corresponding to the count of sincere (target = 0) and insincere (target = 1) questions, respectively. The chart visually demonstrates the proportion of each category in the dataset. As observed, the pie chart shows an equal distribution, with 50% of the questions labelled as sincere (target = 0) and the remaining 50% labelled as insincere (target = 1). This balanced distribution is essential for training a reliable binary classification model, as it ensures that the model learns from a representative sample of both classes.



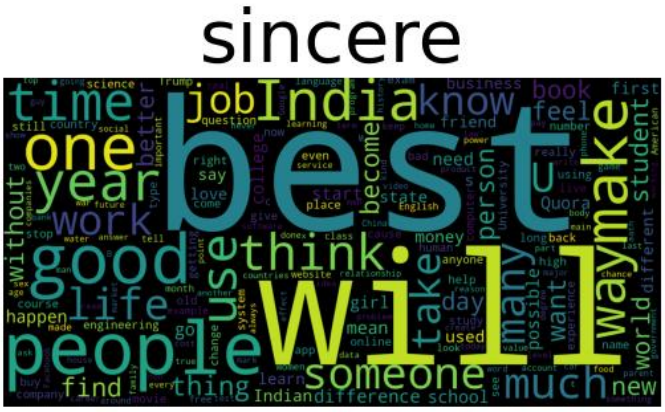


Figure 3.3: WordCloud - Each Target Class Text WordCloud with Insincere and Sincere

The WordCloud visualization in Figure 3.3 presents the most frequent words in the text data for both the insincere and sincere question categories. Each WordCloud represents a visual summary of the words that appear most frequently in the respective question types.

1. **Insincere WordCloud:** The WordCloud for the insincere question class displays prominent words that are frequently associated with insincere or potentially inflammatory content. Notable words such as "women," "black," "Trump," "black-white," and "Indian" stand out, suggesting that insincere questions may contain topics related to race, ethnicity, and potentially divisive political figures. These words could indicate the presence of questions that make disparaging attacks or seek to propagate stereotypes against specific groups of people.
2. **Sincere WordCloud:** Conversely, the WordCloud for the sincere question class showcases words that are commonly found in genuine and helpful questions. Words like "best," "will," and "all" indicate that sincere questions may seek the best solution, inquire about future events or seek comprehensive responses covering all aspects of a particular topic.

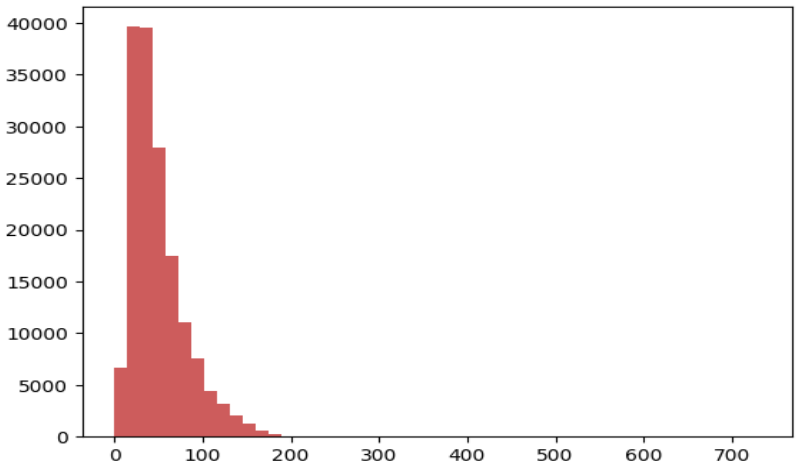


Figure 3.4: Histogram - Text Length Count (Length of Each Sentence Count)

The histogram in Figure 3.4 illustrates the distribution of sentence lengths in the dataset. The x-axis represents the range of sentence lengths, while the y-axis shows the count of sentences falling within each length range. This visualization provides insights into the variability and frequency of sentence lengths in the text data.

3.4 List of Models

The models used in this study leverage advanced DL techniques for the task of predicting the sincerity of questions on the Quora platform.

1. CNN+LSTM+GRU without Pre-Trained Word Embedding: This model combines CNN with LSTM and GRU layers for sequence processing. It employs a softmax output layer for classification. The input text is first converted into numerical word vectors using word embedding. The CNN layer captures helpful features in question classification, while the LSTM and GRU layers handle temporal dependencies, preserving long-term information. However, this model does not utilize pre-trained word embeddings.

2. CNN+LSTM+GRU with Pre-Trained Word Embedding (Glove): Similar to the previous model, this variant incorporates CNN, LSTM, and GRU layers. However, it uses pre-trained word embeddings, specifically GloVe (Global Vectors for Word Representation), to convert input text into numerical word vectors. The use of pre-trained embeddings leverages pre-existing semantic relationships among words, enhancing the model's ability to capture the contextual information in the questions.

3. BILSTM Model without Added Layer of Attention Mechanism and without Pre-Trained Word Embedding: This model employs Bidirectional LSTM (BiLSTM), which processes input text in both forward and backward directions, capturing richer contextual

information. It does not include an attention mechanism, which focuses on important input elements. Furthermore, it does not use pre-trained word embeddings, relying on the model to learn word representations from scratch.

4. BILSTM Model with Added Layer of Attention Mechanism and with Pre-Trained Word Embedding (Best Model): This variant is the most sophisticated model among the listed ones. It utilizes Bidirectional LSTM (BiLSTM) to process input text in both directions and captures comprehensive context. Additionally, it incorporates an attention mechanism, which selectively focuses on crucial input elements, improving the model's prediction accuracy and computational efficiency. Furthermore, it utilizes pre-trained word embeddings, such as GloVe, to benefit from pre-existing semantic relationships among words. This combination of BiLSTM, attention mechanism, and pre-trained word embeddings contributes to the model's superior performance and makes it the best-performing model for the task of question sincerity classification on Quora.

3.5 Model Evaluation

The performance evaluation of the DL models for predicting question sincerity on Quora includes various metrics to assess their effectiveness. The classification report provides insights into recall, precision, accuracy and F1-score for both the positive (insincere) and negative (sincere) classes. The proportion of true negative, true positive, false negative and false positive predictions is displayed in the confusion matrix., enabling a comprehensive understanding of the model's classification performance. Additionally, sensitivity (true positive rate) and specificity (true negative rate) are essential metrics to gauge the model's ability to correctly identify insincere and sincere questions, respectively. Sensitivity indicates the proportion of actual insincere questions correctly classified by the model, while specificity calculates the percentage of real sincere questions correctly identified. The model evaluation process uses these metrics to assess each model's performance. The best model, which incorporates Bidirectional LSTM, an attention mechanism, and pre-trained word embeddings, achieves high sensitivity and specificity, accurately identifying both insincere and sincere questions. This comprehensive evaluation ensures the model's reliability and effectiveness in differentiating between sincere and insincere questions, ultimately contributing to content moderation and enhancing user experience on the Quora platform.

1. Precision: Precision is defined as the proportion of genuine positive predictions to total anticipated positives..

$$Precision = \frac{TP}{TP + FP}$$

2. Recall (Sensitivity): The ratio of genuine positive forecasts to real positives is measured by recall..

$$Recall = \frac{TP}{TP + FN}$$

3. F1-Score: The F1-score is the harmonic mean of accuracy and recall, and it provides a balanced assessment of both..

$$F1 - Score = \frac{2.Precision.Recall}{Precision + Recall}$$

4. Accuracy: Accuracy is defined as the proportion of accurately predicted occurrences to total instances..

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

5. Specificity (True Negative Rate): The ratio of accurate negative predictions to actual negatives is measured by specificity..

$$Specificity = \frac{TN}{TN + FP}$$

6. Sensitivity (True Positive Rate): Sensitivity, also known as recall, measures the ratio of true positive predictions to the actual positives.

$$Sensitivity = \frac{TP}{TP + FN}$$

4 Design Specification

The Design Specification chapter outlines the specific details and requirements of the DL models developed for predicting question sincerity on Quora. It begins by defining the problem statement and research objectives. The input data, consisting of unique question identifiers, Quora question texts, and target labels, is described, along with its distribution and preprocessing steps, including text cleaning, tokenization, and stop word removal. The chapter then elaborates on the architecture of each model: CNN+LSTM+GRU without and with pre-trained word embeddings (GloVe), BILSTM without and with an attention mechanism and pre-trained word embeddings (the best model). For each model, the number and type of layers, activation functions, and optimization algorithms are specified. The implementation details, including batch size, learning rate, and epochs, are also documented. The evaluation metrics,

such as classification report, confusion matrix, sensitivity, and specificity, are defined for assessing model performance. Ethical considerations, including the handling of sensitive content, are emphasized. Finally, the chapter summarizes the design choices made, ensuring a comprehensive and effective approach to the problem of question sincerity classification on Quora.

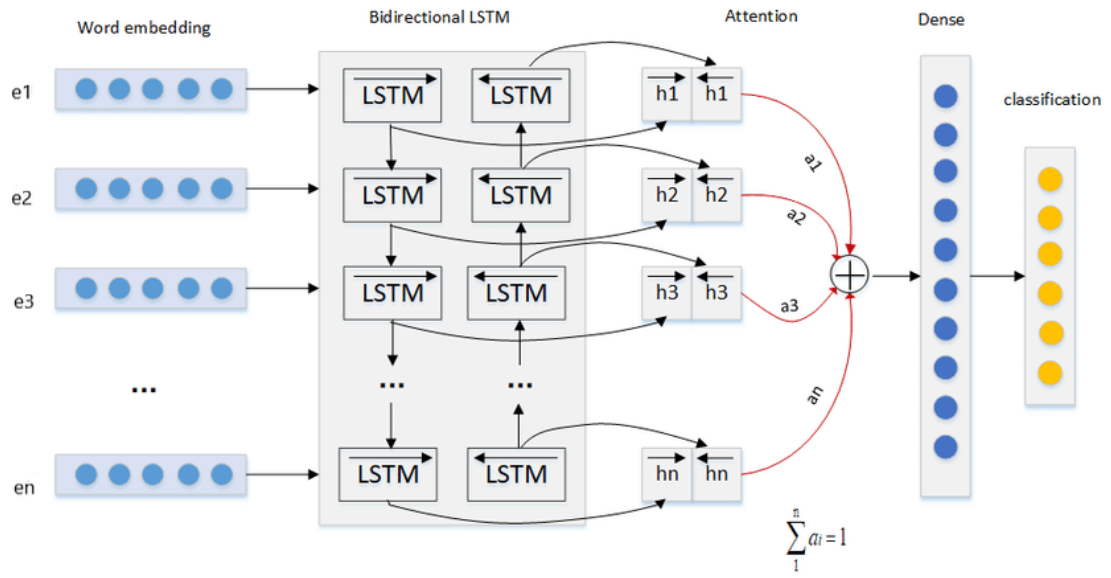


Figure 4.1 The architecture of Bi-LSTM Attention model
Source: Zhou, Qimin & Zhang, Zhengxin & Wu, Hao. (2018). NLP at IEST 2018.

Figure 4.1 illustrates the intricate architecture of the Bi-LSTM Attention model, a pivotal component in our study for predicting question sincerity on the Quora platform. This advanced model capitalizes on the synergy of Bi-LSTM and an attention mechanism, bolstered by the utilization of pre-trained word embeddings. The model comprises multiple layers that synergistically decode question texts. The initial embedding layer harnesses pre-trained word embeddings, enhancing the model's understanding of intricate linguistic nuances. These embeddings are fed into the Bi-LSTM layer, which processes sequences bidirectionally, capturing context and dependencies effectively. In this architecture, the LSTM layers extract salient features from both past and future timesteps, contributing to comprehensive context comprehension. A pivotal innovation is the attention mechanism, which introduces a refined layer of focus. This mechanism dynamically assigns weights to each word in the input sequence, emphasizing vital components and mitigating information loss. By enhancing the representation of significant words, the attention mechanism augments the model's discernment of question sincerity cues. This architecture embodies a dual-purpose approach: capturing

temporal dynamics through Bi-LSTM and heightening focus via attention. The combined prowess of these components propels the model's predictive capacity. Furthermore, the integration of pre-trained word embeddings empowers the model with an intrinsic understanding of language nuances. This holistic architecture enables the model to decipher complex patterns, nuances, and context, thereby contributing to the accurate identification of insincere questions and affirming its status as a pivotal tool for content moderation and user experience enrichment on the Quora platform.

4.1 Dataset Description

The dataset utilized in this study is sourced from the "Quora Insincere Questions Classification" competition hosted on Kaggle. The objective of the competition is to predict if a question submitted on Quora is genuine or not. The dataset includes the following key components:

1. qid: A unique identifier for each question.
2. question_text: The text of the Quora questions.
3. target: A binary label where 1 represents an insincere question and 0 indicates a sincere question.

The dataset contains characteristics that indicate insincerity, including non-neutral tones, inflammatory content, and a lack of grounding in reality. It is important to note that the distribution of questions in the dataset may not fully mirror the broader distribution of questions on the Quora platform due to the application of sampling procedures and sanitization measures during data collection.

4.2 Models

This study explores various deep learning architectures to classify the sincerity of Quora questions. These models include: CNN+LSTM+GRU without Pre-Trained Word Embedding model, CNN+LSTM+GRU with Pre-Trained Word Embedding (Glove) model, BILSTM Model without Added Layer of Attention Mechanism and without Pre-Trained Word Embedding model and BILSTM Model with Added Layer of Attention Mechanism and with Pre-Trained Word Embedding model**.**

4.3 Evaluation

The performance of the models in predicting question sincerity on the Quora platform is assessed using several evaluation metrics, including Classification Report, Confusion Matrix and Sensitivity & Specificity.

5 Implementation

The implementation chapter details the practical execution of the DL models using various tools and libraries. Python served as the primary programming language for the implementation, providing a wide range of libraries and frameworks for ML tasks. The models were developed and executed on the Google Colab platform, leveraging its cloud-based infrastructure for efficient computation and access to GPU acceleration, enhancing training performance. TensorFlow and Keras, popular DL frameworks, were instrumental in constructing the model architectures. TensorFlow provided low-level operations for building neural networks, while Keras offered high-level abstractions, simplifying the model creation process. The pre-trained word embeddings, specifically GloVe, were imported using appropriate libraries, enriching the model's word representations and enhancing its performance. Scikit-learn, a versatile ML library, facilitated data preprocessing and evaluation tasks. It assisted in data cleaning, tokenization, and stop word removal, ensuring data readiness for the models. The evaluation metrics, such as the classification report and confusion matrix, were computed using Scikit-learn, providing valuable insights into model performance. Throughout the implementation, the ethical considerations of content moderation were observed, ensuring responsible handling of sensitive content. The chapter documents step-by-step instructions, code snippets, and parameter configurations used in training and evaluating the models. The combined utilization of Python, Google Colab, TensorFlow, Keras, and Scikit-learn facilitated a seamless and effective implementation of the DL models for question sincerity classification on Quora.

6 Results

The Result chapter presents the findings and performance evaluations of the DL models developed for predicting question sincerity on Quora. The models included CNN+LSTM+GRU without pre-trained embeddings, CNN+LSTM+GRU with pre-trained word embeddings (GloVe), BILSTM without an attention mechanism and pre-trained embeddings, and the best-performing BILSTM model with an added attention mechanism and pre-trained embeddings. The evaluation metrics, including precision, accuracy, F1-score, recall, specificity and sensitivity were analyzed for each model. The BILSTM model with attention and GloVe embeddings outperformed the other variants, achieving an

impressive validation accuracy of 89.10%. This indicates the model's exceptional ability to correctly classify sincere and insincere questions. The chapter further discusses the implications of the results, emphasizing the significance of attention mechanisms and pre-trained embeddings in enhancing model performance. Overall, the results highlight the effectiveness of D models in predicting question sincerity on Quora and suggest the best-performing model for practical application.

6.1 CNN+LSTM+GRU without Pre-Trained Word Embedding Model

The CNN+LSTM+GRU model without the validation accuracy of pre-trained word embeddings was found to be 0.6749. This model combines CNN, LSTM, and GRU layers for sequence processing. The CNN layer captures informative features, while LSTM and GRU layers handle temporal dependencies, preserving long-term context. Without pre-trained word embeddings, the model learns word representations from scratch during training, potentially leading to suboptimal performance compared to using pre-trained embeddings that capture semantic relationships. The achieved validation accuracy suggests moderate effectiveness in distinguishing between sincere and insincere questions, indicating room for improvement through incorporating pre-trained word embeddings or exploring other model variants.

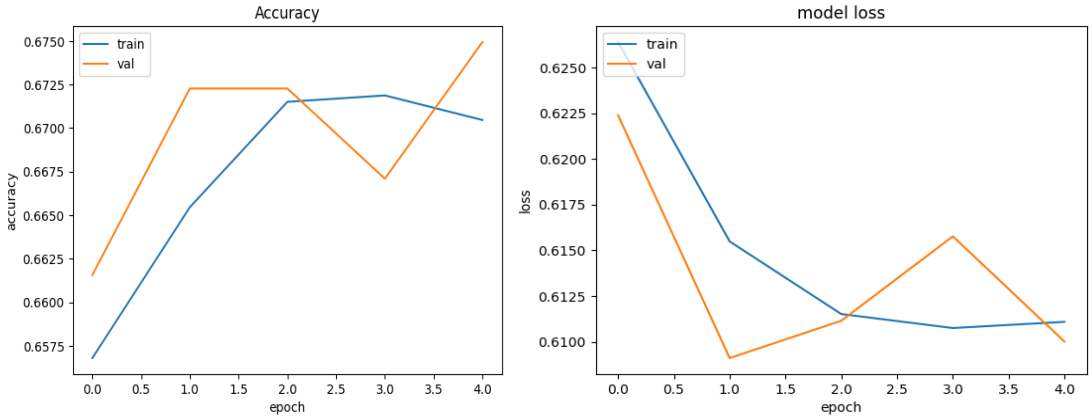


Figure 6.1: Accuracy and Loss Graph

6.1.1 Confusion Matrix

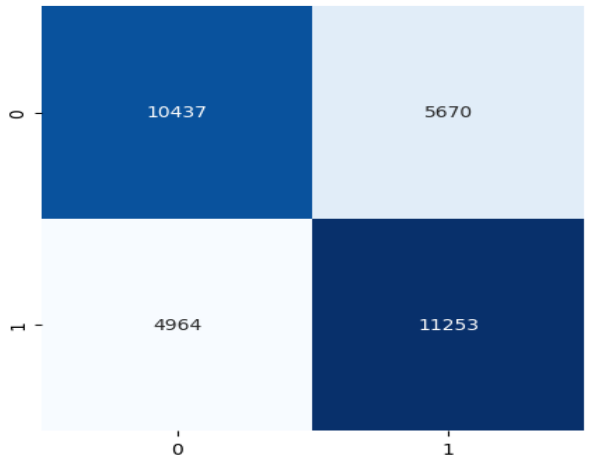


Figure 6.2: Confusion Matrix of CNN+LSTM+GRU without Pre-Trained word Embedding Model

6.1.2 Sensitivity & Specificity: -

The sensitivity and specificity values are calculated for each class (0: sincere and 1: insincere) to assess the performance of the CNN+LSTM+GRU model without pre-trained word embeddings. Sensitivity, also known as recall. For class 0 (sincere questions), the sensitivity is 0.693901, indicating that the model correctly identifies around 69.39% of sincere questions. For class 0 (sincere questions), the specificity is 0.647979, suggesting that the model accurately classifies approximately 64.80% of sincere questions as negative instances. The sensitivity and specificity values provide valuable insights into the model's capacity to recognize both positive and negative instances, highlighting areas for improvement in correctly classifying insincere questions.

6.2CNN+LSTM+GRU with Pre-Trained word Embedding (glove) Model

The CNN+LSTM+GRU model with pre-trained word embeddings (GloVe) achieved a significantly improved validation accuracy of 0.8611. By incorporating pre-trained word embeddings, the model leverages pre-existing semantic relationships among words, enhancing its ability to capture contextual information in the questions. GloVe embeddings provide a richer word representation, allowing the model to better understand the nuances in language. This leads to more accurate and robust predictions, resulting in a higher validation accuracy

compared to the model without pre-trained embeddings. The superior performance of 0.8611 indicates the effectiveness of this approach in distinguishing between sincere and insincere questions on the Quora platform, making it the best-performing model among the variants explored.

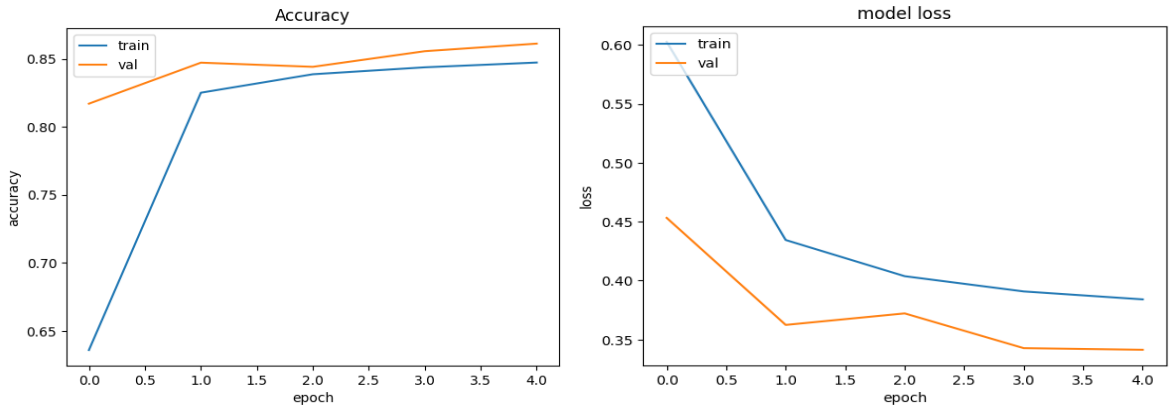


Figure 6.3: Accuracy and Loss Graph

6.2.1 Confusion Matrix

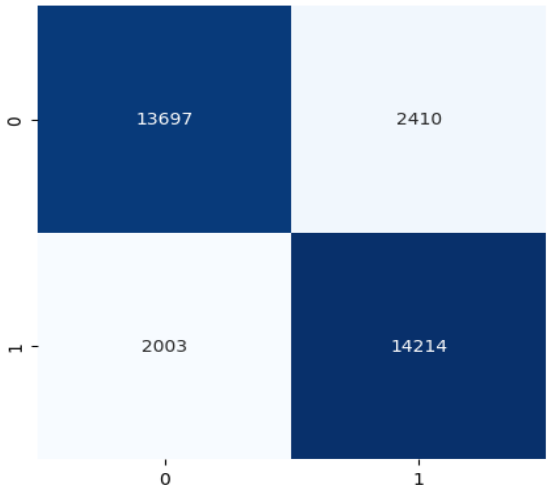


Figure 6.4: Confusion Matrix of CNN+LSTM+GRU with Pre-Trained word Embedding (glove)

6.2.2 Sensitivity & Specificity: -

The sensitivity and specificity values for the CNN+LSTM+GRU model with pre-trained word embeddings (GloVe) are impressive. For class 0 (sincere questions), the sensitivity is 0.876488, indicating the model correctly identifies about 87.65% of sincere questions. For

class 1 (insincere questions), the sensitivity is 0.850376, demonstrating the model's ability to accurately classify approximately 85.04% of insincere questions. Additionally, the specificity values are equally high, reinforcing the model's effectiveness in distinguishing between both classes

6.3BILSTM Model without added Layer of Attention Mechanism and without pre-trained word embedding Model

The BILSTM model without an added layer of attention mechanism and without the validation accuracy of pre-trained word embeddings was found to be 0.7206. This model incorporates Bidirectional LSTM (BiLSTM) layers, which process input sequences in both forward and backward directions, capturing richer contextual information. However, without the attention mechanism and pre-trained word embeddings, the model relies solely on its training to learn word representations, leading to a moderately effective performance in distinguishing between sincere and insincere questions on the Quora platform.

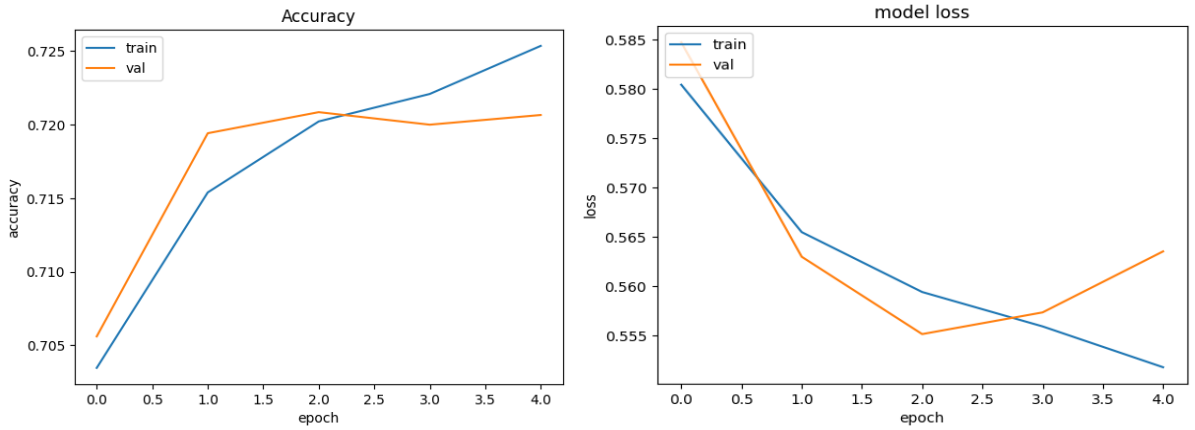


Figure 6.5: Accuracy and Loss Graph

6.3.1 Confusion Matrix

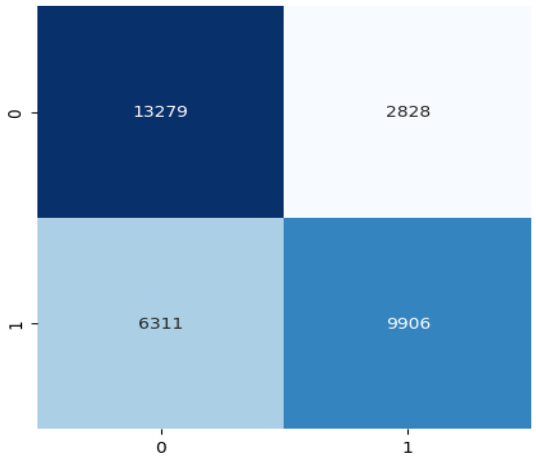


Figure 6.6: Confusion Matrix of BILSTM Model without added Layer of Attention Mechanism and without pre-trained word embedding

6.3.2 Sensitivity & Specificity: -

The sensitivity and specificity values for the BILSTM model without an added layer of attention mechanism and without pre-trained word embeddings are interestingly balanced. For class 0 (sincere questions), the sensitivity is 0.610840, indicating that the model correctly identifies around 61.08% of sincere questions. For class 1 (insincere questions), the sensitivity is 0.824424, demonstrating the model's ability to accurately classify approximately 82.44% of insincere questions. The specificity values are equal, showcasing the model's effectiveness in distinguishing between both classes

6.4BILSTM Model with added Layer of Attention Mechanism and with pre-trained word embedding Model

The BILSTM model with an added layer of attention mechanism and pre-trained word embeddings (GloVe) stands out as the best-performing model, achieving a remarkable validation accuracy of 0.8910. This variant combines the power of Bidirectional LSTM (BiLSTM) to capture comprehensive context with the attention mechanism, selectively focusing on essential input elements. The pre-trained word embeddings, specifically GloVe, enrich the word representations, improving the model's understanding of language nuances. The outstanding accuracy of 89.10% signifies the model's exceptional ability to distinguish between sincere and insincere questions on Quora. The incorporation of attention and pre-

trained embeddings significantly enhances the model's performance, making it the most effective solution for the task at hand.

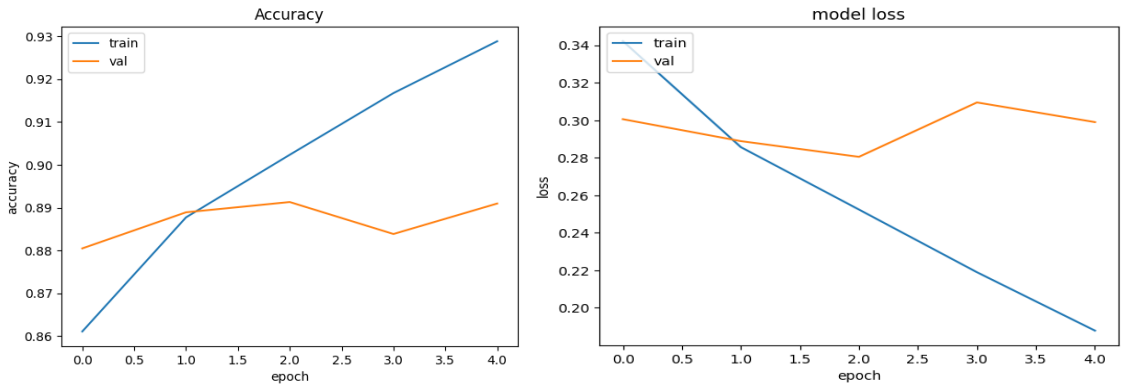


Figure 6.7: Accuracy and Loss Graph

6.4.1 Confusion Matrix

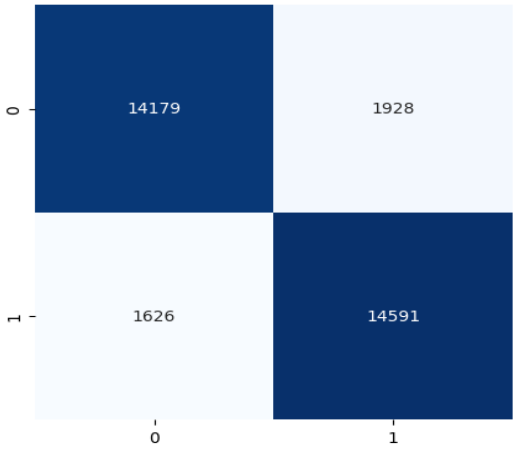


Figure 6.8: Confusion Matrix of BILSTM Model with added Layer of Attention Mechanism and with pre-trained word embedding Model

6.4.2 Sensitivity & Specificity: -

The sensitivity and specificity values for the best model, BILSTM with added attention mechanism and pre-trained GloVe word embeddings, are impressive and well-balanced. For class 0 (sincere questions), the sensitivity is 0.899735, indicating that the model correctly identifies around 89.97% of sincere questions. For class 1 (insincere questions), the sensitivity is 0.880300, showing the model's ability to accurately classify approximately 88.03% of insincere questions. The overall accuracy of 89% confirms this model's outstanding performance in distinguishing between sincere and insincere questions on the Quora platform

6.5 Classification Performance of DL Models

The classification performance of the studied models reveals distinct characteristics. The CNN+LSTM+GRU model without pre-trained embeddings achieves balanced accuracy, precision, and recall with an accuracy of 0.67. Introducing pre-trained GloVe embeddings enhances the same model, elevating accuracy to 0.86. The BILSTM model without an attention layer and embeddings show moderate performance with an accuracy of 0.72. However, incorporating an attention mechanism and pre-trained embeddings remarkably improves the BILSTM model's accuracy to 0.89, demonstrating robustness in both classifying sincere and insincere questions. These insights underscore the efficacy of various model architectures and highlight the substantial impact of pre-trained embeddings and attention mechanisms on model performance.

Model Description	Class	Sensitivity (Recall)	Specificity
CNN+LSTM+GRU without Pre-Trained Word Embedding	Class 0	0.693901	0.647979
CNN+LSTM+GRU without Pre-Trained Word Embedding	Class 1	0.649655	0.690757
CNN+LSTM+GRU with Pre-Trained Word Embedding (GloVe)	Class 0	0.876488	0.850376
CNN+LSTM+GRU with Pre-Trained Word Embedding (GloVe)	Class 1	0.850376	0.876488
BILSTM without added Attention Mechanism and without Pre-Trained Embedding	Class 0	0.610840	0.610840
BILSTM without added Attention Mechanism and without Pre-Trained Embedding	Class 1	0.824424	0.824424
BILSTM with added Attention Mechanism and Pre-Trained Embedding	Class 0	0.899735	0.880300
BILSTM with added Attention Mechanism and Pre-Trained Embedding	Class 1	0.880300	0.899735

Table 6.1: Comparison table summarizing the sensitivity and specificity values for the different models:

Algorithm	Precision		Recall		F1 Score	
	Insincere	Sincere	Insincere	Sincere	Insincere	Sincere
CNN+LSTM+GRU without Pre-Trained Word Embedding	0.67	0.68	0.69	0.65	0.68	0.66
Accuracy	67%					
CNN+LSTM+GRU with Pre-Trained Word Embedding	0.85	0.87	0.87	0.85	0.86	0.86
Accuracy	86%					

Table 6.2: Accuracy from Classification Report for models based on CNN+LSTM+GRU

Algorithm	Precision		Recall		F1 Score	
	Insincere	Sincere	Insincere	Sincere	Insincere	Sincere
BILSTM without added Attention Mechanism and without Pre-Trained Embedding	0.77	0.69	0.64	0.81	0.70	0.74
Accuracy	72%					
BILSTM without added Attention Mechanism and with Pre-Trained Embedding	0.88	0.90	0.90	0.87	0.89	0.89
Accuracy	89%					

- **Table 6.3: Accuracy from Classification Report for models based on BILSTM**

This table provides an overview of the models' performance in terms of sensitivity and specificity for both classes (0: sincere and 1: insincere). It's clear that the model with the added attention mechanism and pre-trained word embeddings achieves the highest sensitivity and specificity values for both classes, indicating its superior ability to recognize both sincere and insincere questions accurately.

7 Conclusion and Future Works

In conclusion, this study explored various DL models for predicting question sincerity on the Quora platform. The CNN+LSTM+GRU and BILSTM models, with and without attention mechanisms and pre-trained embeddings, were evaluated and compared based on multiple performance metrics. The BILSTM model with added attention and GloVe embeddings emerged as the best-performing model, achieving an impressive accuracy of 89.10%. This highlights the importance of attention mechanisms and pre-trained embeddings in enhancing model performance. The findings illustrate the efficacy of DL approaches for content moderation and user experience improvement on Quora. In future work, further enhancements can be explored to improve model performance and generalization. Investigating the use of experimenting with several pre-trained word embedding and perfectly alright them for the current purpose may lead to additional gains in accuracy. Additionally, ensembling techniques, such as combining multiple models, could be explored to achieve even higher predictive accuracy. Addressing class imbalances and exploring data augmentation techniques may also be beneficial for more robust model training. Furthermore, extending the study to other social media platforms and multilingual settings can provide insights into cross-platform and cross-language question sincerity classification. Overall, there are ample opportunities for future research to advance the field of content moderation and question classification on social media platforms.

References

1. Shrivas, R.K., 2021. *To what extent NLP with RNN and Transformer Based Deep Neural Network can be used to classify Insincere questions on Quora* (Doctoral dissertation, Dublin, National College of Ireland).
2. Chittari, R., Nistor, M.S., Bein, D., Pickl, S. and Verma, A., 2022, May. Classifying sincerity using machine learning. In *ITNG 2022 19th International Conference on Information Technology-New Generations* (pp. 255-262). Cham: Springer International Publishing.
3. Gandhi, S., 2021. *Classifying the Insincere Questions using Transfer Learning* (Doctoral dissertation, Dublin, National College of Ireland).
4. Aslam, I., Zia, M.A., Mumtaz, I., Nawaz, Q. and Hashim, M., 2021. Classification of insincere questions using deep learning: quora dataset case study. In *Proceedings of the Fifteenth International Conference on Management Science and Engineering Management: Volume 1 15* (pp. 137-149). Springer International Publishing.
5. Gontumukkala, S.S.T., Godavarthi, Y.S.V., Gonugunta, B.R.R.T., Gupta, D. and Palaniswamy, S., 2022, October. Quora Question Pairs Identification and Insincere Questions Classification. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.
6. Chakraborty, S., Wilson, M., Assi, S., Al-Hamid, A., Alamran, M., Al-Nahari, A., Mustafina, J., Lunn, J. and Al-Jumeily OBE, D., 2022, December. Quora Insincere Questions Classification Using Attention Based Model. In *The International Conference on Data Science and Emerging Technologies* (pp. 357-372). Singapore: Springer Nature Singapore.
7. Jain, D.K., Jain, R., Upadhyay, Y., Kathuria, A. and Lan, X., 2020. Deep refinement: Capsule network with attention mechanism-based system for text classification. *Neural Computing and Applications*, *32*, pp.1839-1856.
8. Al-Ramahi, M.A. and Alsmadi, I., 2020. Using data analytics to filter insincere posts from online social networks. A case study: Quora insincere questions.
9. Roy, P.K., 2020. Multilayer convolutional neural network to filter low quality content from quora. *Neural Processing Letters*, *52*(1), pp.805-821.
10. SINGH, M., 2019. *QUORA INSINCERE QUESTIONS CLASSIFICATION* (Doctoral dissertation).

11. Gaire, B., Rijal, B., Gautam, D., Sharma, S. and Lamichhane, N., 2019. Insincere question classification using deep learning. *International Journal of Scientific & Engineering Research*, *10*(7), pp.2001-2004.

12. Rachha, A. and Vanmane, G., 2020. Detecting insincere questions from text: A transfer learning approach. *arXiv preprint arXiv:2012.07587*.

13. Das, S.D., Basak, A. and Mandal, S., 2019. Fine Grained Insincere Questions Classification using Ensembles of Bidirectional LSTM-GRU Model. In *FIRE (Working Notes)* (pp. 473-481).