Unethical and Bad Questions Detection by Using Machine Learning Techniques

Lokesh Naidu Bavigadda
LokeshNaiduBavigadda@my.unt.edu

Vamsi Krishna Chittimadha
Vamsikrishnachittimadha@my.unt.edu

: *Sri Sai Bhargava Kakarlapudi*
srisaibhargavaseetharamarajukakarlapudi@my.unt.edu

l M Manikanta Venkata Pasumarthi
pasumarthiraghuram12@gmail.com

github : LINK

l***Abstract*—**

An insincere question aims to make a statement rather than seeking genuine information. It often employs a non-neutral, exaggerated, or rhetorical tone to convey a particular point about a group of people. These questions may be disparaging, discriminatory, or based on false premises. The process of training a machine learning classifier to identify insincere questions involves analyzing a dataset of labelled questions and evaluating its performance on a test set. This analysis demonstrates the classifier's effectiveness in distinguishing insincere questions. To enhance the effectiveness of identifying insincere questions, three machine learning algorithms were trained and compared: MultinomialNB, BernoulliNB, and LogisticRegression. Each algorithm was evaluated on a separate test set to assess its ability to accurately classify insincere questions. The results indicated that Logistic Regression outperformed the other two algorithms, demonstrating its superior ability to distinguish genuine inquiries from insincere statements

***Keywords*—MultinomialNB, BernoulliNB, Logistic Regression**

Motivation and Significance:

The process of training a machine learning classifier to identify insincere questions involves analysing a dataset of labelled questions and evaluating its performance on a test set. This analysis demonstrates the classifier's effectiveness in distinguishing insincere questions. To enhance the effectiveness of identifying insincere questions

Objectives:

Automate the detectioning of insincere questions, harmful content with Machine Learning Models for responsible and safe internet.

.

Features:

the dataset consists of Non neutral tone It might have an exaggerated tone to understand a point from a group of people. May be a statement from a group of people. There may be a isn't grounded in reality. The train data consists of insincere and sincere questions labelled as 1 and 0.

I. Introduction

Insincere questions, often posed with malicious intent, can significantly hinder constructive dialogue and create a hostile environment for users. To address this challenge, I utilized the data provided in the Kaggle competition and employed advanced machine learning techniques to effectively distinguish between genuine inquiries and insincere statements. The research process encompassed several crucial steps, including data cleaning, feature engineering, and model selection. By meticulously preparing and analyzing the data, I was able to train and evaluate various machine learning models, ultimately selecting the most effective models for the task at hand. The implementation of these machine learning models represents a valuable contribution to the ongoing efforts to create a safer and more respectful online environment. By automating the detection of insincere questions, harmful content can be effectively filtered out, fostering a more positive and productive user experience.

I. DataSet

Data has 130612 unique questions where there are two labels insecure and secure questions. The dataset was collected from the Kaggle website which is a famous and trusted platform for getting data. Whereas test data has 37508 unique rows. This has enough numbers to trust fot the test data. Where the dataset consists of Non neutral tone It might have an exaggerated tone to understand a point from a group of people. May be a statement from a group of people. There may be a isn't grounded in reality. The train data consists of insincere and sincere questions labelled as 1 and 0.

II. Detail design of features

The dataset underwent a comprehensive preprocessing pipeline, encompassing data cleaning, punctuation removal, contraction replacement, case normalization, negative word replacement, acronym handling, stop word removal, elongated word identification, stemming, lemmatization, and TF-IDF vectorization. These techniques collectively refined the data, enhancing its quality and suitability for downstream analysis. Lemmatization is a normalization technique used in Natural Language processing to group different forms of words so that they can be analysed in a single item. Stemming is more advanced which removes the ending of the word to get it from the root. TF-IDF vectorization measures the frequency of the word in a document The higher TF value indicates the word appears more frequently in the document. Stop words that do not add meaning to the text are removed and the process helped to focus the analysis to get the more meaningful out of our dataset. To eliminate redundancy and improve data consistency, elongated words, which are words with repetitive characters, were replaced with their standard counterparts. This approach ensured that each word was represented in its most concise form, facilitating more efficient analysis and interpretation. And handled capitalised words in the dataset for the design of features. The dataset was loaded into the pandas data frame for a better handle of the dataset with the optimised memory..

II.Analysis

The graph (Fig 1.0) below shows a comparison of the average number of syllables per word in insincere and sincere questions. The graph shows that the average number of syllables per word in insincere questions is higher than the average number of syllables per word in sincere questions. This means that insincere questions tend to have longer words than sincere questions.There are a few possible explanations for this difference. One possibility is that insincere questions are more likely to be complex and require more sophisticated language. Another possibility is that insincere questions are more likely to include rare or technical words. The graph also shows that there is a greater range of average syllables per word for insincere questions. This means that there is more variability in the complexity of insincere question. The median average syllables per word for sincere questions is 15, while the median average syllables per word for insincere questions is 20. This suggests that the average insincere question is significantly longer than the average sincere question.The standard deviation for the average syllables per word is higher for insincere questions than for sincere questions. This suggests that there is more variability in the complexity of insincere questions

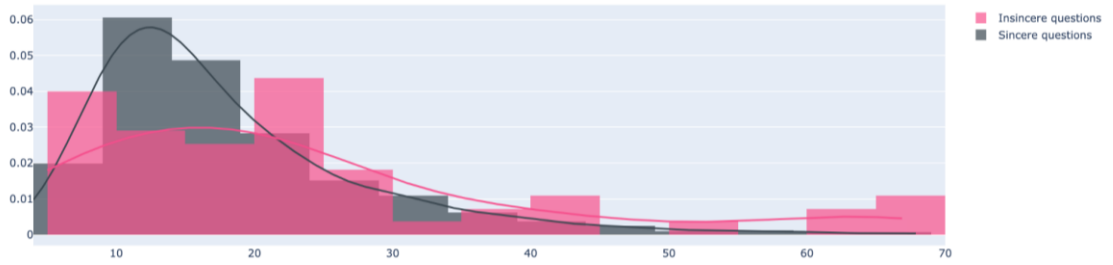

Figure 1 Syllable Analysis

The figure two illustrates a positive association of question length and the mean syllables every words. Therefore, we can say that the higher the number of words in a question, the greater is the average number of syllables per word in that question. Nonetheless, such a correlation works better with insincere than sincere questions which implies that insincere queries are much more complicated to involve in sophisticated speech. In fact, such findings suggest that complex and sophisticated words are more likely to occur in insincere questions as well. Besides, the number of average syllable per word appears to reach higher levels with longer questions. Therefore, it implies that the questions become more complex and varied as they increase in length. Sincere and insincere questions are both covered by this observation; however, it is more pertinent to insincere questions.This implies that different types of questionnaires like long questions can be employed in all sorts of studies among them sincere and non-sincere queries. This can exemplify a lengthy question that seeks elaborate clarification, complex assignment or a query meant to provoke thought. Again, this graph can be used for deceiving somebody, manipulating someone, and just making jokes. Finally, it shows that the average number of syllables is likely to determine the intensity of questioning. An average number of syllables is more indicative for dishonesty in questions. Nevertheless, there are different variables that may also affect the genuineness of the question like the time and place when the question was raised or if the person who asked the question was in a good mood or not so sincere.

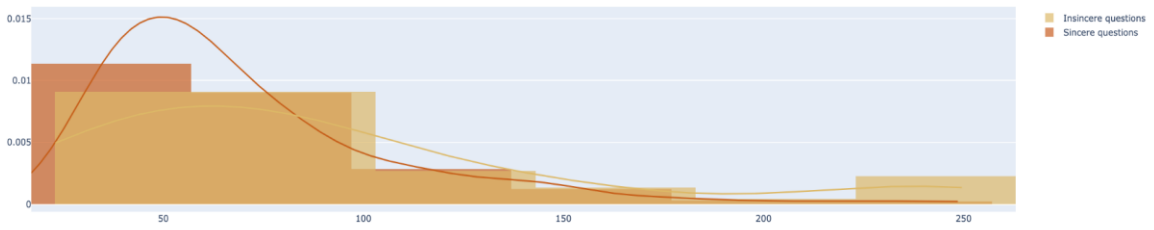

Figure 2 Question Length

Figure 2 demonstrates that question length is positively correlated with average syllables per word. As a result, when questions are long, the average number of syllables per word in a question also becomes higher. Nevertheless, this connection is more obvious for insincere than sincere queries, which indicates about complexity of language in case of insincere questions. This implies that, probably, inauthentic queries incorporate more

sophisticated vocabulary or less common words and phrases.Furthermore, it reveals higher scores for averagely syllabic per words within extended posed questions. The line in this graph indicates the association of question lengths with mean syllables per word in both sincere and insincere queries. It implies that as a question lenghtens, the average number of syllabes per word within the question also rises. The strength of the correlation increases however, if it's about insincere questions than sincere questions. This implies that complex questions and sophisticated language can also arise from an inauthentic line of questioning. This indicates that untrue questions will mostly include few and specialized words.In addition, it brings out wider spectrum of average syllable per word for the sincere and insincere questions. In other words, longer questions are much more variable with regard to their complexity.

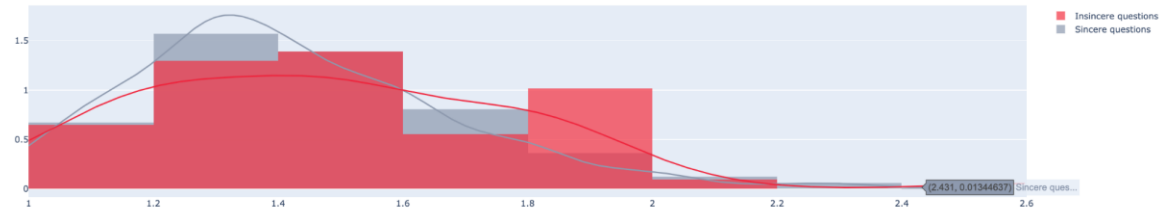

Figure 3 Average syllables per word

The graph Figure 4 says that that there is a positive relationship between question length and mean syllables per word for both genuine and fake questions. Thus, the amount of syllables on word basis for an average will grow with increasing length of a question. Nonetheless, it is stronger in case of insincere questions as compared to the sincere ones.It implies that dishonest questions would probably be intricate and utilize complicated words. Furthermore, it implies that false inquiries are likely to contain uncommon and technical terms.It is also worthy to note that the average syllables per word span is wider for longer questions in case sincere as well as inauthentic questions. In other words, longer questions differ much in their complexity.This means that longer questions can serve for different situations with both true or false motives. For instance, questions like how much details does the person want to know about? To what extent is the given task difficult? How can someone's judgment be challenged? are such kinds of questions. It could also be employed to cheat somebody, control someone or simply for jesting purposes.Generally, this plot indicates that the average number of syllables per word may serve as an appropriate measure of the sincerity of the questions, whereby the questions having a higher average number of syllables are more likely insincere ones

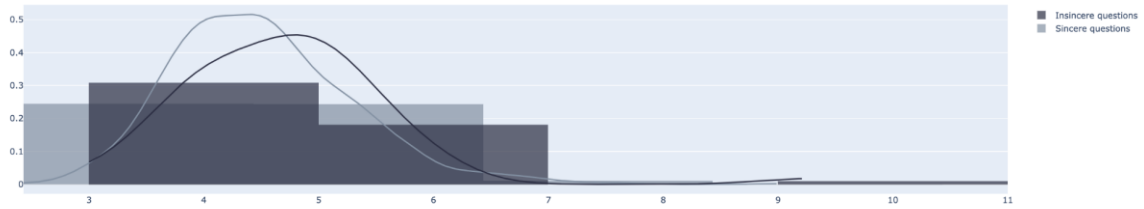

Figure 4 Average letters per word

Flesh Reading score is a redability test it was developed in 1940. It is based on two factors average length of sentence and average length of syllable per word. It is used to tell how easy to read a text. The formula goes like this FRES = 206.875 - (0.846 * ASL) - (1.015 * ASW). ASL is an average sentence length. ASW is an average syllabus per word. If the score is 100 to 90 it is very easy to read. If the score is 89 to 79 it is easy to read. If the score is 69 to 50 fairly easy to read. For 49 to 30, it is fairly difficult to read. 29 to 20 difficult to read. 19 to 0 very difficult to read. So we tested this on our dataset to check the redabality of our questions in our dataset the results shows like this.

The graph Figure 5 shows that majority of the insure questions has FRES score has less than 70 which is a little difficult to use. That means insure questions are difficult to read. And it also says that more range of FRES score for insecure questions . that means there is more readability for insecure questions. The median FRES score for insincere questions is 58, and the median FRES score for sincere questions is 73. It means that average insecure questions than a secure questions.

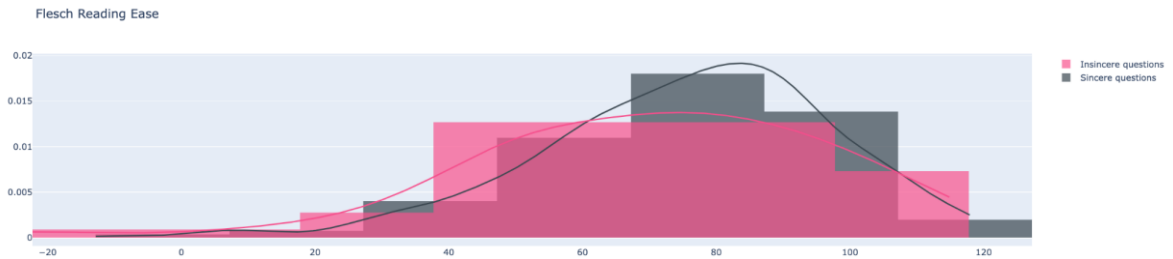

Figure 5 Flesch Reading Ease

I. Implementations

The Dataset we are using is a text based dataset. So we performed data cleaning operations for a better model prediction we did some pre processing steps such as clean text, clean numbers, miss spell dictionaries, removed stop words replaced constactions, lemen text and cleaning sentence, these are the function names that are coded in python for preprocessing steps, which are mentioned in ipynb file cells. And we loaded the data in data frame format to perform pre processing steps. And we also did tfidf_vectorizer it is a famous vectorizer concept in nlp techniques that is used to get better model results for text based datasets. Here we used sklean library for a builtin function of tfid vectorzation and count vectorozation.

We used kfold validation helps to prevent overfitting and easy to reliable for a best model. So here we used 5 splits in k fold validation and trained one algorithm called logistic regression. Logistic regression is a statical model which is also known as logit model. It is used for classification and predictive model. It estimates the probability of an event occured. For an example vote or dint vote based of an independent variable. So the probability of an outcome lies between 0 and 1 in logistic regression a logit transofmrations is applied to the odds one. That is probabilty of the sucess divided by the probabilty of the failure. This is also commonly known as log odds. Or the natural odds algorithm, and the formulas goes like this

$$\text{Logit}(pi) = 1/(1+ \exp(-pi))$$

$$\ln(pi/(1-pi)) = \text{Beta}_0 + \text{Beta}_1*X_1 + \ldots + B_k*K_k$$

In this regression equation, logi - pi is independent to the responsible or respsonbl variable and x is the independent variable. The beta parameter is likely estimated via maximum likely hood estimation that is MLE. through multiple iterations this method etsts different values fo the best fir of logs log likehoods function. Our optimal coefficient will be found. Conditional probabilities for every observation is calculated and logged to yield a predicted probability. To find the binary classification a probability less than 0.5 will be preditexed probability.

The log odds are difficult to interpret with a logistic regression analysis . but as a result to beta estimates is common to transform results into a odd ration that is easy to interpretation of result. The or represents the odds that will occusring in the absence of that event. IF the OR is greater then 1 te event is associated to the higher odds that to generate a specific outcome.

They are three types of logistic regression they are Binary logistic regression, Multinomial logistic regression. Ordinal logistic regression.

In Binary logistic regression. The dependent variable is independent in nature it has only two outcomes. That means a binary outcome . For example, it used to predict wether the email is spam or not or whether is to hack or not.this is the vdely used example to explain lgostic regression. In Multinomial logistic regression. There are three outcomes. Example to understand what type of drink consumers prefer based on location and the age. Or three based regression. The last one is ordinal logistic regression. This logistic regression model is used to response variable that has more than three outputs. These values do have order. For example movies rating that can be between 0 to 5.

The other examples where logistic regression is used other than this dataset IT can be used to find fraud detection, disease prediction or churn prediction.

II. Preliminary Results

As mentioned above the results for the Logistic regression model. The metric used here is the F1 score from sklearn library it is used to evaluate the model performance. It tells how many times our model a correct prediction for our entire dataset It used precisi on and recall scores of the model and computes it

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

F1 score for our model in the test data. We got 0.59. In Increment 2 we are going to explore different models to get more result than the present result.

III. Project Management

Implementation status report:

- Work completed
  - Description
    - As mentioned above the data was collected from Kaggle which is open sourced data. Performed traditional techniques, which are to be done sequentially. First, we did
      - data cleaning.
        - Removed numbers from our dataset in a required column which is question column
        - Replaced repetition of punctuations with respective sting as required for our dataset.
        - Removed punctuations for the same column.
        - Replaced contractions for the question column
        - Lowered the case for the entire daset
        - Replaced negations with antonyms as mentioned in the code
        - And also handled capitalised words
        - Removed stop words which are largely occurred in our dataset.
        - Replaced elogngagted words in the question column
        - And performed stemming and lemiataions
        - TFI vectorization which is mandatory for out logistics regression algorithm for a better result.
      - Analysis
        - Counted the total syllable analysis
        - The lexicon analysis
        - Analysed the question length for question column
        - Avergage syllabus per word in q question
        - Average letters per word in a question
        - Flesh reading ease formula which gave very insight meaning for our dataset
  - Responsibility

Task	Person	Status	Contribution
Data Cleaning	Lokesh	completed	25%
Analysis	Vamsi	completed	25%
Feature Generation	Bhargava	completed	25%
Model Training	Manikanta	completed	25%

- Work to be completed
  - Description
    - Has a scope in advanced data cleaning where the model can easily converge for a better results
    - Looking for more analytical methods and statical methods for the good understanding of the dataset. Where a lot of change will be there for model selection which has to bt trained
    - Going to extract more features from the dataset.
    - Machine Learning models such as MultinomialNB and BernoulliNB will be trained
    - Hyperparameter tuning should be done for the models for a better features for an algorithm
  - Responsibility
    - Models should be developed without biased information
  - Issues/concerns
    - As dealing with a text based dataset with 45MB in size running on a local machine takes times and has computation limitations.

References

1. Rajadesingan, A., Liu, H., & Nourbakhsh, A. (2015). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 24th International Conference on World Wide Web (pp. 1371-1376).https://www.researchgate.net/publication/262348492_Detecting_offensive_tweets_via_topical_feature_discovery_over_a_large_scale_twitter_corpus
2. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web (pp. 145-153). https://dl.acm.org/doi/10.1145/2872427.2883062
3. Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex machina: Personal attacks seen at scale. Medium. Read Article https://arxiv.org/abs/1610.08914

4. Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.https://aclanthology.org/D14-1162/
5. Depression detection using emotional artificial intelligence and machine learning https://www.sciencedirect.com/science/article/abs/pii/S2214785322005430
6. A review on sentiment analysis and emotion detection from text Pansy Nandwani1 · Rupali Verma https://rdcu.be/dpujk
7. Analyzing quora for the insinceres https://www.kaggle.com/code/thebrownviking20/analyzing-quora-for-the-insinceres
8. Text pre processing techniques https://www.kaggle.com/code/deffro/text-pre-processing-techniques
9. Naive Bayes and logistic regression baseline https://www.kaggle.com/code/stardust0/naive-bayes-and-logistic-regression-baseline
10. Feature-engineering-for-nlp-classification https://www.kaggle.com/code/shaz13/feature-engineering-for-nlp-classification
11. Conventional-methods-for-quora-classification https://www.kaggle.com/code/mlwhiz/conventional-methods-for-quora-classification
12.