# Insincere Questions Classification

## Introduction:

An existential problem for any major website today is how to handle toxic and divisive content. Internet trolls act by posting divisive content on websites to provoke controversy and emotional reactions from readers. While less scrupulous sites may not mind the added traffic trolling provides, question and answer forums like Quora that focus on providing credible information must take steps to moderate these types of posts. They cannot allow their platform to be used for provocative questions or confirm hateful stereotypes. These questions, also termed as insincere questions, are intended to make a statement rather than look for helpful answers.

This project will focus on detecting insincere questions posted by members on the Quora QA community forums using advanced machine learning techniques. Insincere questions typically have political, non-neutral tone, or convey an extreme view of a group. Training machine learning models to detect such features is not a straightforward endeavor. While topics of conversation may provide a clue for a question's intentions, features for clearly dividing "troll" type questions from those seeking real answers are not obvious and depend on the context of the sentence in its entirety.

## Why is the problem important?

Statista estimates that there are 4.65 billion social media users as of April 2022, roughly 58% of the global population. Social media sites are now a reality of life and show no signs of disappearing in the near future. Unfortunately, some users choose to engage with social media in ways that violate social norms or the content policies of the sites. In order to avoid negative press, or civil or criminal liability, social media operators must moderate the content hosted on their sites. As more companies and organizations add social media capabilities to their online presence, the amount of data generated by their users grows vast and it becomes impractical, if not impossible, to review all of this data manually. In order to lighten the load of moderators, these groups have turned to machine learning to automate the task, and identify potentially disparaging or inflammatory content so that it can be reviewed before being displayed to a broad audience. In this paper, we will propose a method to improve upon existing strategies to identify social media content posted on the Quora website that warrant further review, with the ultimate goal that this strategy could be extended to use by other social media platforms.

## Literature Review

As pointed out by Mungekar et al., in their research regarding insincere question classification, one of the most important parts in text classification and mining is the preprocessing. The results of the different text classification evaluation indicators show that the TF-IDF algorithm has certain advantages in text classification.

Another such paper proposed a model for classifying tweets (Indra, 2016). The authors used a Logistic Regression model in order to classify tweets according to the topic. The model first transformed the tweets into vector, which is similar to the one mentioned about by another paper (Fan, 2018)]. The model used the word vector to calculate accuracy. The confusion matrix showed an accuracy of around 92%. A wide variety of classification techniques have been used to document classification [ (Rossi, 2016)]. The models developed include Naive Bayes, Logistic Regression, Support Vector Machine (SVM), an ensemble of Naive Bayes and Logistic Regression and Random Forest. While all of the models provided high accuracy rates, the F1 score and ROC provided more meaningful model performance metrics due the data being unbalanced. (Agarwal, n.d.).

In another research, (Dong, n.d.) it focuses on the automatic keyword-based operations carried out in terms of keyword indexing, classification, clustering along with five different keyword extraction methods. In addition, 2-way ANOVA has been used to validate the performed analysis. The study states that the use of ensemble approach consisting of Bagging based Random Forest method provided the accuracy around 93%. Various SVM and Naive Bayes approaches are used and compared and a later stage in the paper. The author concludes by stating that this approach offers performance and computational efficiency advantages versus conventional methods.

Support Vector Machine Model was used to classify BBC documents into five categories (Haryanto, n.d.). The model was enhanced by using Chi-Squared along with SVM. Both Stemming(Lancaster Stemmer) and Lemmatization (WordNet Lemmatizer) were used and fed separately to the model. The results stated that Stemming provided better results and chi-squared was an added advantage as it improved the model.

Text classification was used for identifying tweets related to suicides (Chiroma, n.d.). This was done with the motive of reducing the negative impact of tweets. The models used for this project included SVM, Naive Bayes and Random Forest. Decision tree performed the best in among the three models implemented. The F-measure ranged from 0.346 to 0.778.

## Proposed Approach

**Data Pre-processing**:

The first step in any NLP (Natural Language Processing) based application is to cleanse the dataset by correcting spelling and grammar mistakes and remove inconsequential or redundant words. The Dataset has only two elements – question text and classification whether sincere or not. We plan to extract a few additional metadata from the question text and add them in the dataset

- n_words = Number of words in Question
- numeric_count = Number of numeric words in Question
- special_character_count = Number of special characters in Question
- unique_words = Number of unique words in Question
- char_words = Number of characters in Question
- count_misspelled_word – count of incorrectly spelled words in the questions

These additional features may help us evaluate the data better in feature extraction.

**Word Vs Sentence Embedding:**

Word embedding is the language modeling technique in natural language processing where individual words or phrases are represented as a real-valued vector that can capture the context of the word in a document, semantic and syntactic similarities, and relation with each other word. The same concept applied to sentences within a dataset is Sentence embedding. This paper will use both techniques and compare the performance when used in different modeling types.

**Modeling**:

The embeddings will then be used in supervised machine learning algorithms and neural networks. Logistic regression will be one of the techniques used and is best known to be used in a problem with binary outcome. Neural networks are formed of hierarchical multistage layers. The first layer is sentence embedding which maps each sentence to a vector space. The second layer can utilize either CNN, RNN, or LSTM to get a contextual representation of the question. Subsequent layers can focus on feature extraction and the last layer will determine the question classification.

**Training and Validation:**

The dataset is quite large and may not need cross validation. An 80-20 split will be used to tune the models. For the Neural networks, the epoch size, and batch size will be tuned by multiple runs.

**Extension:**

As an extension, we can perform word frequency analysis and keyword extraction to attempt to categorize the dataset by topic, and the frequency of words or phrases commonly used in insincere versus sincere comments. A model can be created for each of the question categories. This will require more research to create sentence embeddings specific to a topic.

## Proposed method for evaluation

Due to data imbalance the evaluation is not focused on Accuracy, rather it is focused on other metrics like F1 score, Area Under Curve, Precision and Recall. These metrics are explained below.

1) Accuracy: Accuracy can be said to the measure of the closeness of the output to a certain value. It does not work well with imbalanced data. Hence, in this project other metrics are used for evaluation.

2) Precision: Precision is ratio of correctly predicted outcomes to the total predicted outcomes. It is not dependent on the accuracy of the model. It is therefore a measure that be used in case of class imbalance.

3) Recall : It is the ratio of correctly predicted outcomes to the total outcomes. It is also known as the sensitivity of the model.

4) F1 Score: F1 score is the one that is calculated by combining the precision and recall measures. It is the harmonic mean of the two. It results nearly the same as the average of the two measures when they are closely related.

Various evaluation metrics will be considered as the data is highly imbalanced. The accuracy level cannot be considered for judging the best model as even if all questions are to be considered sincere the accuracy will be above 90%. Hence, F1 score acts as the main metric for evaluating the performance of the models.

## Time plan of the project

| Project Stages | Dates |
|---|---|
| **Project Proposal** | 6/15/2022 |
| **Exploratory data analysis** | 6/22/2022 |
| **Feature Engineering/NLP API Research** | 6/29/2022 |
| **Project Milestone 1 documentation** | 7/06/2022 |
| **Modeling** | 7/13/2022 |
| **Training and Classification** | 7/20/2022 |
| **Performance Analysis** | 7/27/2022 |
| **Final Report/ Project Milestone 2** | 8/03/2022 |
| **Final Presentation** | 8/10/2022 |

## References (APA formatting in final paper)

1. https://www.kaggle.com/competitions/quora-insincere-questions-classification
2. https://huggingface.co/blog/1b-sentence-embeddings
3. https://en.wikipedia.org/wiki/Quora
4. https://www.researchgate.net/publication/334549103_Quora_Insincere_Questions_Classification (Mungekar, n.d.)
5. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26647500.pdf (Mao, n.d.)
6. https://link.springer.com/chapter/10.1007/978-3-030-79203-9_12 (Aslam, 2021)
7. https://ieeexplore.ieee.org/document/8492945 (Liu, n.d.)
8. https://medium.datadriveninvestor.com/approaching-the-quora-insincere-question-classification-problem-eb27b0ad3100 (Mohanty, 2020)
9. https://www.atlantis-press.com/proceedings/ncce-18/25896557 (Fan, 2018)
10. https://www.researchgate.net/publication/314667612_Using_logistic_regression_method_to_classify_tweets_into_the_selected_topics (Indra, 2016)
11. https://ieeexplore.ieee.org/document/1358033 (Dong, n.d.)
12. https://ieeexplore.ieee.org/abstract/document/8549748 (Haryanto, n.d.)
13. https://ieeexplore.ieee.org/document/8527039 (Chiroma, n.d.)
14. https://www.sciencedirect.com/science/article/abs/pii/S0306457315000990?via%3Dihub Google Scholar (Rossi, 2016)
15. https://link.springer.com/chapter/10.1007/978-1-4614-3223-4_6 (Agarwal, n.d.)
16. https://www.statista.com/statistics/617136/digital-population-worldwide/