# Classification of Insincere Questions using SGD Optimization and SVM Classifiers

Akshaya Ranganathan, Haritha Ananthakrishnan, Thenmozhi D, Chandrabose Aravindan

Department of CSE, SSN College of Engineering, Chennai
{akshaya16009,haritha16038}@cse.ssn.edu.in {theni_d, aravindanc}@ssn.edu.in

**Abstract.** The burgeoning of online question and answer forums like Quora, Stack Overflow, etc. has helped answer millions of questions across the globe. However, this also simultaneously engendered the problem of Insincere Questions. Insincere questions are those that have a nonneutral undertone with no intention of seeking useful answers. The Classification of Insincere Questions task at FIRE 2019 did not just focus on binary classification of insincere questions, but went a step further and introduced a fine-grained classification of 6 labels, namely: rhetorical, hate speech, sexual content, hypothetical, other and sincere questions. The solution offered by SSN_NLP used a 2 level approach. The fundamental level comprised of SGD optimized SVM Classification model. In addition to this, the corpus was filtered based on frequently occurring, relevant keywords. This approach produced an accuracy of 48%

**Keywords:** Classification · Insincere Questions · SVM Classifier · SGD Optimizer · Relevant Keywords

[1]

## 1 Introduction

With millions of questions being asked daily across myriad Q & A platforms, there is a pressing need to eliminate redundant and potentially dangerous insincere questions. The sheer volume of questions asked every single day makes it a mammoth task for human moderators. Question forums such as Quora, Stack Overflow and Yahoo Answers used to deploy methods to manually weed out such insincere questions before the advent of Machine Learning and Natural Language Processing. If left unchecked, these questions can cause serious issues to the platform as well as the general morale of the users. Such insincere questions may even lead to a drastic decrease in the number of users or in some cases,

or can negatively impact several users. Thus identification of Insincere questions is cardinal. The data provided is an enhanced subset of the dataset provided by Quora. The Quora dataset consists of 2 labels. The CIQ dataset consists of 6 labels of questions: - Rhetoric, Hate Speech, Sexually objectionable, Hypothetical, Sincere and other questions. This fine-grained classification aims at not just better identification of such questions but also providing fitting countermeasures for each type of question. SSN_NLP proposed a supervised learning approach. The solution is twofold: On the root level is SGD optimized SVM Classifier after vectorizing. This solution is further refined by filtering using the most relevant keywords of the Hate Speech and Sexually objectionable content categories. The overall accuracy of the model is around 53%.

## 2 Related Works

Classification of insincere questions was originally an online competition hosted by Kaggle: a platform for data scientists to implement Machine Learning concepts and engage in a world-wide competition with other data scientists, powered by Quora [5]. This task mandated the use of word embedding models such as GloVe and Word2Vec for text vectorization and had over 1400 submissions. **Cross-industry standard process for data mining** (CRISP -DM) was used by one of the teams in [1], which is a life cycle model of knowledge discovery, such as KDD. They used a combination of Logistic Regression and Naive Bayes classifiers, support vector machines, Decision Trees and Random Forrest algorithms for their modeling. A Deep Learning approach with Gated Recurrent Units (GRU) was deployed by [2] in addition to supervised learning methodologies. Apart from question forums, mass social networking sites like Twitter and Facebook are being monitored for posts that are ironical, or pose as any kind of security threat [6]. Our work involves the use of SGD Optimization and TF-IDF Vectorization, with an additional layer of relevant keyword extraction mechanisms.

## 3 Dataset Analysis and Preprocessing

The training data released by FIRE 2019's CIQ task included 899 user's questions that were assigned label belonging to 6 target values, whose distribution is given in Table.1 The main challenge of data preprocessing was removing stopwords and converting the text to lower case for uniformity.

- **T**ext **tokenization** was performed using NLTK [2] toolkit, to divide the sentence into words, which were later compared with a stop-word list. NLTK English stopword list was used for this purpose.
- In addition to this, **punctuation marks, special characters, and numbers** were removed using regular expressions, as a consequence of text tokenization.

[2] https://www.nltk.org/

**Table 1.** Distribution of training data

| Label | Description | Count |
|---|---|---|
| 0 | Sincere Questions | 20 |
| 1 | Rhetorical Questions | 488 |
| 2 | Sexual Content | 98 |
| 3 | Hate Speech | 216 |
| 4 | Hypothetical Questions | 38 |
| 5 | Other | 38 |

- Since online questions contain numerous amounts of **contractions** such as can't, won't, etc. (See Fig .1) and **misspelled words** such as "*beleived*" we wanted to eliminate the fallacious effect they would have on the overall accuracy if they were considered as separate words. Contracted words were expanded to their actual meaning, and misspelled words were corrected using pre-defined lists.
- The **Porter-Stemmer** stemming algorithm was used to link any particular word to its root. For example, the words **speaking** and **speaks** both are derived from the root word **speak**. Hence, it would be much more practical for the model to understand any derivatives of a root word, as the root itself to evade discrepancies relating to the weight vectors of words.

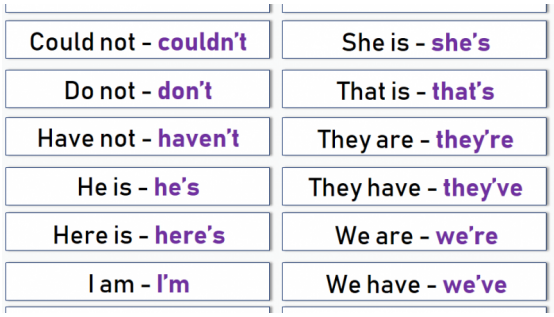


**Fig. 1.** Common English Word contractions

## 4 Methodology and Implementation

The model used TF-IDF vectorizer [3]. Term Frequency (TF) implies the ratio of the number of times a word is occurring in a document to the total number of words in the document. It gives a measure of the frequency of a term in a document. However, words like ***the*** will have a higher term frequency, but they have little importance in the document. This problem is solved by using the

Inverse Document Frequency (IDF). IDF gives a measure of the occurrence of a term across all documents of the corpus. The model uses Stochastic Gradient Descent Optimiser upon Support Vector Machine Classifier (SVM) [4]. SGD was the preferred choice of optimizer as it works relatively efficiently even in large datasets, as the gradient of the loss is estimated one sample at a time, which is randomly shuffled for performing the iteration. For a use case like the classification of Insincere questions, this would prove efficient and scalable. Further, a relevant keyword identification layer was added that scanned the training data and extracted unique words that were not in the stop word list. Using sklearn's **Count Vectorizer**, frequently occurring keywords for the labels Hate Speech, and Sexual Content was obtained. Keywords for classes such as Hypothetical and Rhetoric questions could not be easily distinguished from other English words, and thus were not taken into consideration to avoid False positives.

## 5 Result Analysis

| Actual/ Predicted | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| 1 | 2 | 35 | 5 | 18 | 4 | 3 |
| 2 | 0 | 3 | 6 | 1 | 2 | 1 |
| 3 | 1 | 4 | 3 | 6 | 0 | 0 |
| 4 | 0 | 0 | 0 | 2 | 1 | 0 |
| 5 | 0 | 2 | 0 | 0 | 0 | 0 |

**Fig. 2.** Predicted Vs. Actual Label distribution

The dataset contained 899 rows and 6 labels. The distribution of data amongst the labels was widely varied ranging from about 488 questions for the Rhetoric question category to a meager 20 questions in the Sincere Question category. This distribution was reflected in the models performance. The model performed with an accuracy of about 47.52%. A closer analysis of the results indicates that the model performed well in identifying rhetoric questions. However, it was not able to identify questions in the *Sincere* and *other* question category. These categories had the least training data share. Moreover, the model also used a second layer

of filtering based on relevant keywords for the hate speech and sexually objectionable question categories. The overall accuracy of the model can be improved by having a more evenly distributed dataset. Some categories had extremely few questions which proved any kind of re-sampling method futile. Yet another direction to focus on while improving accuracy is to build efficacious lists of keywords for each category. The distribution of predicted labels for our model is given in Fig. 2, where the diagonal elements are the True Positives.

## 6 Conclusion and future works

Classification of insincere questions is a concept that will benefit not only social media Q&A platforms but also its users. It helps ensure the virtual safety of its users and creates a neutral environment. It is also a cost-effective, and time-efficient method, as it involves limited manpower as opposed to extant methods of content quality assurance. Our team, SSN_NLP used a supervised learning model and produced an accuracy of 47.52%. Future developments for our model would include
1. Using a pre-trained word embedding model such as Google's BERT to improve our F1 scores.
2. Applying semi-supervised learning techniques to deal better with smaller datasets

## References


1. Mungekar, Akshay, Nikita Parab, Prateek Nima, and Sanchit Pereira. "Quora Insincere Questions Classification."
2. Gaire B, Rijal B, Gautam D, Sharma S, Lamichhane N. Insincere Question Classification Using Deep Learning.
3. Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications.181. 10.5120/ijca2018917395
4. Robbins H, Monro S. A stochastic approximation method. The annals of mathematical statistics. 1951 Sep 1:400-7.
5. Kaggle Inc. 2019. Quora Insincere Questions Classification: Detect toxic content to improve online conversations. `https://www.kaggle.com/c/quora-insincere-questions-classification`
6. Deshwal A, Sharma SK. Twitter sentiment analysis using various classification algorithms. In2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) 2016 Sep 7 (pp. 251-257). IEEE.