Quora Insincere Questions Classification

Arnold Namegni
University of Bremen
Bremen, Germany
djomonam@uni-bremen.de

Kyounghyun Bae
University of Bremen
Bremen, Germany
kbae@uni-bremen.de

Mosharrafa Ahmad
University of Bremen
Bremen, Germany
mos_ahm@uni-bremen.de

ABSTRACT

Quora has been one of popular question-answer platforms and functioned as a place where users share their knowledge with other people around the world. Since it is an online website and anonymity still exists, there are some questions which contain toxic that provokes certain hatred or insults or contain insincerity that are not intended to be answered. This project was carried out to build a machine learning model to predict whether a question asked on Quora is sincere or insincere. We proposed several approaches to solve this machine learning problem:Binary classification, bag of words and Watson Tone Analyzer. After cleaning and tokenizing through lemmatization, we applied 3 different models to train our dataset which consists of labelled questions(insincere question:1, sincere question:0): Logistic regression, Support vector machine(SVM) and multi-layer perceptrons(Neural Networks). The best result was found in logistic regression with an accuracy of 83% during the cross validation and 84% during the prediction on the test value. Unfortunately, application of Watson Tone Analyzer did not give a valuable result for the prediction.

KEYWORDS

Data science, Machine learning, Natural language processing, Binary classification, Logistic regression, Quora

1 INTRODUCTION

Quora is a website where users of websites voluntarily post questions and answer each other's questions. It has started its service from 2010, and until now it has grown into one of the largest, and popular question-answer based websites and it has been used across the world. People can post questions from various topics, follow a question, follow a topic, share questions and its answers [7, p. 1].

Quora has a policy that requires users to use their real names for the access and use of the website. Users can also register to the website through Facebook or Google. However, an actual verification of a user's name is not mandatory, and as the platform has grown so fast, there is a necessity to handle qualities of questions. Since it is a user-based social Q&A website, anonymity still plays an important role and this actually propose a potential that some malicious, toxic questions can be posted. These questions threatens the value of free social Q&A websites and destroys Quora's ecosystem of knowledge sharing.

Our project aims to detect these toxic, and insincere questions which intends to provoke certain types of hatred, insults or does not intend to realistic answers. In this project, we work on constructing a machine learning model to predict insincere questions and make this platform biased-free and valuable place where people can share their knowledge freely worldwide.

2 RELATED WORK

Quora has been a popular theme for various scientific research papers. Among previous studies, there are two research papers that are based on data science and analyzed the questions on Quora. One paper analyzes the question topic popularity in Quora and predict the next popular topic of questions[7]. It examines dynamics of topic growth and understand various key factors associated with popularity of topics [7, p. 2]. The other study analyzes the linguistic structure of question texts, characterizes and predicts the answerability of question in Quora [6].

Our study is different from previous studies, in the aspect of the purpose of the study. Indeed, this project is specifically focused on ethics, morality and the public interest so that everyone can benefit from the sharing and acquisition of knowledge freely and safely without the risk of being insulted, attacked and confused.

3 BACKGROUND

There are several machine learning concepts that are used to approach this study. These concepts are derived from various machine learning books and previous data science researches. To build our machine learning model, we employed a program language *Python* and a software program *Jupyter Notebook*.

Supervised Learning: Binary Classification

Classification is one of the most common supervised machine learning problem, whose goal is to predict a *class label* from a predefined list of possibilities. [4, p. 27]. 3 different machine learning algorithms are selected to use for the machine learning model. One is *logistic regression* which is commonly used in text classification, especially in binary classification. Second is *linear support vector machine* which is also one of the popular algorithm for classification, especially suitable for a smaller, medium-sized dataset through maximizing the margin between two classes. Third one is multi-layer perceptrons which can build up more complex model which is sensitive to parameters. [4, p. 106].

Natural Language Processing (NLP): Bag of words

Natural Language Processing is a process of transforming texts into something more suitable for Machine learning algorithm. There are several methods for NLP. However, for this project, *bag of words* is employed to extract features from the words used in questions to train the classifier models for this project. Each word is considered as a feature. To make feature extractions more clearly, *lemmatization* is employed, which is an advanced tokenization method that represent each Word with its word *stem.* [4, p. 346]. The declaration of this function is built in a Python library called *Spacy*, which is a library for advanced Natural Language Processing.

Watson Tone Analyzer

In the aforementioned study which analyzes the linguistic structure of question texts, characterizes and predicts the answerability of question in Quora, it says that a questioner's linguistic, emotional, cognitive states are revealed through the language he/she uses in the question text [6, p. 613]. From this idea, it is necessary to examine the relationship between tone and the questions and we thought there might be a correlation between tone and the insincerity of questions.The research made by IBM on texts shows that human writings can also contain the tone of their authors [2]. Based on this research, IBM developed this tool **Tone Analyzer** which can identify the tone of a text within 7 cateogories: Tentative, confident, analytical, sadness, joy, fear, anger

Accuracy, Precision, Recall and f-score

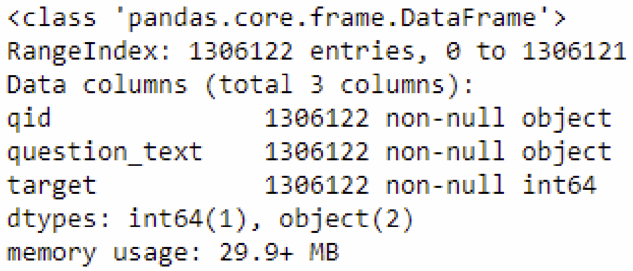
There are 4 metrics that are used to evaluate machine learning algorithms. While *accuracy* provides us the information about the percentage of good prediction, *precision* measures how many of samples predicted as positive are actually turned out to be positive. On the other hand, *recall* measures how many of the positive samples are captured by the positive predictions. This is important since this project aims to avoid false negatives (Insincere questions predicted as sincere questions)[4, p. 285]. Through these metrics, a validation of machine learning model is decided for this project.

4 METHOD

Dataset Description

The data we used in this paper was obtained from Kaggle [5], which is one of the most popular online community for data scientists. Kaggle allows data scientists to find public datasets and also hosts data science and machine learning competitions [1].The dataset proposed on Kaggle have approximately 1,31 millions of (figure 1) questions which are labelled either 0, or 1. 0 means the question is sincere, on the other hand 1 means the question is insincere.

There are four main criteria that imply the insincerity of questions [5].First is **non-neutrality of tone**: If a question exaggerates or underscores a point about a group of people, it is insincere [5]. Second,if a question is **disparaging or inflammatory** that intends discrimination, stereotyping or insults, it denotes insincerity [5]. Third, an insincere question is **not based on reality or a logical assumption**[5]. Lastly an insincere question uses **sexual contents** for shock value or attention [5].



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1306122 entries, 0 to 1306121
Data columns (total 3 columns):
qid              1306122 non-null object
question_text    1306122 non-null object
target           1306122 non-null int64
dtypes: int64(1), object(2)
memory usage: 29.9+ MB
```

Figure 1: Dataset basics informations.

Machine learning model

Data pre-processing and Feature extraction: As shown in figure 2, the dataset is heavily *imbalanced.* Therefore, the first step was to balance it by randomly selecting the same number of questions for both classes: 80810 questions from sincere questions and 80810 questions from insincere questions. To reduce computational costs of algorithms, we will work on only 5% of data from the balanced dataset, which is exactly 4040 Question. We also divided this last into two part, 80% for the training phase and 20% for the test dataset.

Furthermore, machine learning algorithms cannot work directly on the text. To apply the machine learning algorithms, each question was transformed into a vector(Bag of words) using lemmatization as tokenization method. All the *stopwords* [1] i., punctuation marks were deleted and all the letters turned into lower cases.

Through Watson Tone Analyzer, the tone of each question was identified and added at the end as a new feature named **ton**. The tone of some questions couldn't be defined by the tool. Therefore, we labelled those question with *undefined* and used *FeatureUnion* to combine the result of those two steps (tokenization and Tone Analyzer) together .

[1] Word that are too frequent in natural text to be informative

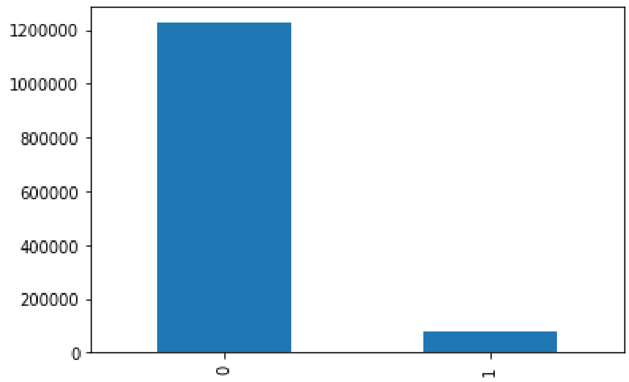


Figure 2: Label distribution.

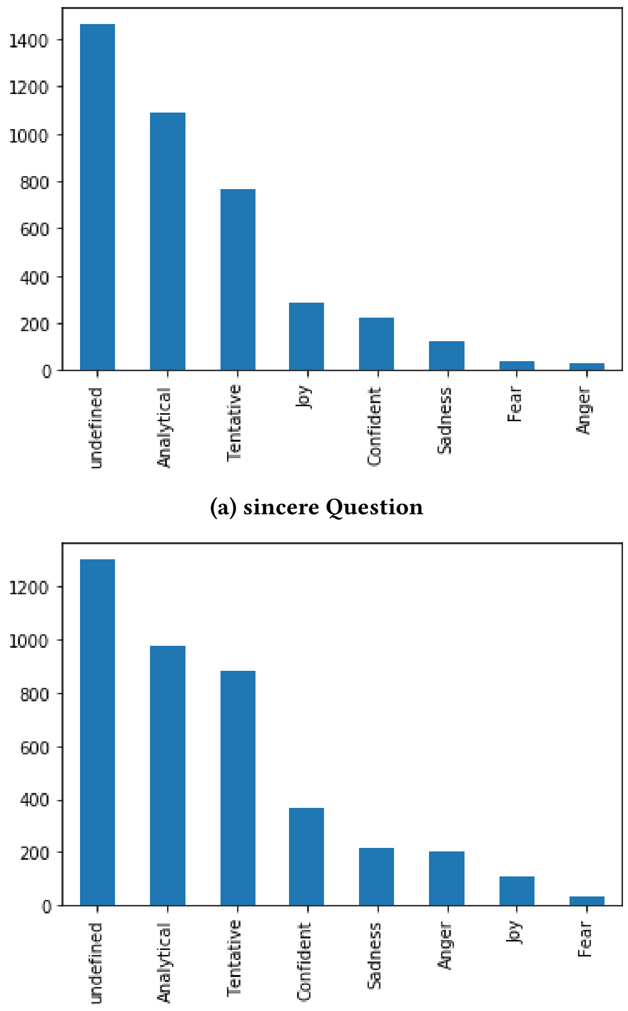
Model Selection. : We then trained and compared several machine learning models that work best on text classification such as: Logistic regression, SVM (Support Vector Machine) and Multilayer perceptrons for classification (Neural Networks). To do so, we firstly performed a 5-fold cross-validation with the dataset only containing the tokenized questions and secondly with the dataset containing the tone features.

Tuning and Prediction. : In this last part, the selected models will be trained with the best parameter for the prediction on the test dataset.

5 PERFORMANCE OF MODELS

Figure 3 shows the tones of the different questions featured from Watson Tone Analyzer. Unfortunately, there are a considerable amount of questions whose tones are undefined. Moreover, 3 does not show any relationship between the tones of questions and insincerity. This is due to the fact that most of them have been identified as either tentative or analytical. We then was a little bit skeptical at this point that this feature will have a negligible effect on our prediction model or no effect at all.

Table 1 summarizes GridsearchCV combined with a 5-fold Cross-validation made on the different algorithms. As 1 shows, the different results of the different algorithms seem to be very similar. However the model which employed the logistic regression without tone analyzer offers the best results with an accuracy of 83% during the cross-validation and 84% during the prediction on the test value. Additionally, we trained the K-nearest Neighbor classifier on our dataset but we didn't expected to have a good result, since it's not suitable for our data type. However, since its implementation is simple, we tried this model only to evaluate the amount of words that should be necessary for the decision. Therefore, we trained it with an unigram tokenization and searched the best number of neighbor, which was 5. We concluded therefore that the range of values of the parameter *ngram_range* for the tokenization function would be between 1 and 5, and this reduced the search interval for the gridsearchCV on other algorithms.



(a) sincere Question

(b) insincere Question

Figure 3: Tone Captured on the dataset

To train the model with the tone of the questions we had to remove from the dataset all the samples with undefined tone. Unfortunately as we can see on table1, this tone feature does not help to improve the prediction of model.

Moreover, gridsearchCV allowed us to choose our classifier as logistic regression with the following parameters: $C = 10$,and $ngram\_range = (1, 3)$. The chart (Figure4) shows the 30 most important features for the prediction of each class made by logistic regression model. The height of the bars show the the value of each coefficient and therefore

Table 1: Cross-Validation Dashboard

Algorithm	Best Accuracy	Accuracy on Test
KNN	0.69	0.70
Logistic regression (LR)	0.83	0.84
LR with tone feature	0.81	0.84
SVM	0.83	0.83
Neural network	0.81	0.84

its importance. The positive coefficient on the right are the group of words that predict a insincere question. The left we have a group of words which predict a sincere question. Some *tri-gram* like “Chinese woman ugly” are surprisingly classified among the decisive words for a sincere question. This is not actually a mistake, since the question containing those word was labelled in the datatset as sincere. On the other hand they are quite intuitive feature like : “racist” ,“hate”,“pussy” other “gay”, which clearly implies that the content of the Question will be about a group of person. It is also very interesting to note that all the 20 first decisive features for the prediction of Question insincere are composed entirely of unigrams. This is probably due to the selection of dataset.

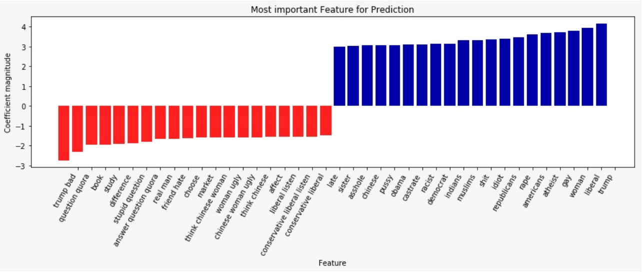
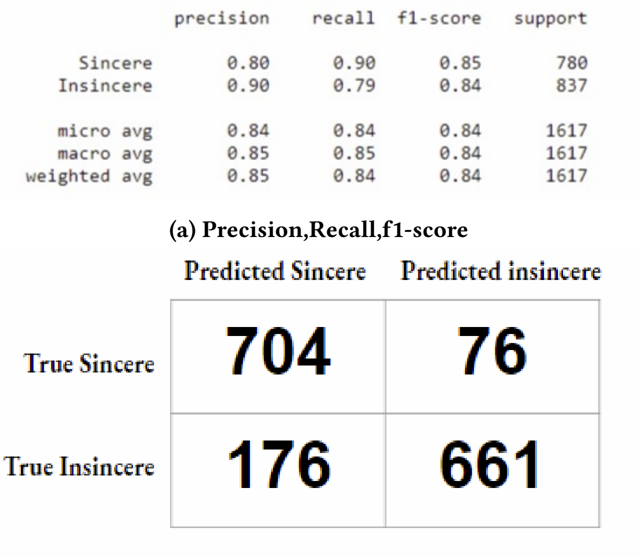


Figure 4: 20 Most important Feature of Logistic regression.

On the figure 5 and we can have a look on the classification report of logistic regression. Within 780 sincere questions, 704 are predicted as sincere questions.However, the most important thing to see is the prediction made on insincere question. Within 837 insincere questions, the model made 79% good prediction. This is actually the value of the recall on insincere questions.This value give us a significant information about the amount of false negatives made in the prediction, since for our project, it is important to reduce the number of false negatives, i.e. the number of insincere questions predicted as a sincere question.

6 LIMITATION

We tried to approach this project with the 3 models which are well-know for solving classification problems. Additionally, we wanted to figure out if a tone of a question has a significant impact on the prediction of the insincerity of questions. However, we would like to address difficulties and limitations of this project regarding to the selected models and tones of questions.



```
             precision    recall  f1-score   support

    Sincere       0.80      0.90      0.85       780
  Insincere       0.90      0.79      0.84       837

  micro avg       0.84      0.84      0.84      1617
  macro avg       0.85      0.85      0.84      1617
weighted avg       0.85      0.84      0.84      1617
```

(a) Precision,Recall,f1-score

	Predicted Sincere	Predicted insincere
True Sincere	704	76
True Insincere	176	661

(b) Confusion matrix.

Figure 5: Classification report of the Logistic regression

Limited selection of models

For this project, three machine learning models were employed: Logistic regression, Linear support vector machine (SVM), Multi-layer perceptrons. Logistic regression gave us the best prediction, however there is a lack of tryouts from other machine learning models regarding to text classification. All those mentioned three machine learning models are simple to use and known for classifications. We could have tried more basic models of classification, such as naive Bayes or decision trees. However, there could a better way to solve our machine learning problem with a more sophisticated approach. For example, *word2vec* could be our machine learning problem solution. Word2vec can group the vectors of similar words together in a vectorspace and allow to detect the similarities among the words and give features about the context of individual words. Through this model, we might have a better result for the prediction of our test dataset.

Pitfall of Watson Tone Analyzer

Watson Tone Analyzer, developed by IBM, is a tool to detect an emotional tone of a person from his or her text. IBM has built its own ensemble framework to set up the 7 tone categories: Anger, Fear, Joy, Sadness, Analytical, Confident and Tentative. It is based on the theory of psycholinguistics, which understands whether the words that people use in their day-to-day lives reflect who they are, how they feel, and how they think [3]. Their dataset was collected through twitter customer-support forums as the source of conversational data [3].

Therefore, there is a difficulty to apply this tool to our study since our dataset consists of only questions. Moreover, the contexts of datasets between Watson Tone Analyzer and our study are different: IBM used the texts from the area of customer service, and our dataset came from a social QA websites. This might be the reason why a lot of questions from our dataset turned out to have a undefined tone.

Another pitfall is that we falsely assumed that the words used in the questions can represent certain types of emotions. However, through detailed observations of our dataset, we found out that the words employed in questions do not explicitly express emotions, rather certain words imply some types of prejudices or hatred. For example, one of insincere questions from our dataset says "Why is Quora dominated by right-wingers?". It does not represent any kind of emotion, but it contains some kind of hatred about right-winged people with the use of word "dominated". There is another question that says "Why don't more gay men attempt to decrease their feelings of same-sex attraction?". The words in this question does not express any emotion, however it contains a prejudice against homosexuals, like if the love of homosexuals is bad. Therefore, we should have employed another tool to catch the prejudice or hatred behind the words or catch the style the way of a questioner uses words, so that we could have improved our prediction more precisely.

7 CONCLUSION

Our study is aimed at classifying the questions from Quora, a social Q&A website, into two categories: sincere or insincere. To build a machine learning model for the prediction of insincere questions, we used the approaches of binary classification and natural language processing. For our binary classification, we employed 3 algorithms: Logistic regression, linear support vector machine and multi-layer perceptrons. Before applying these algorithms, we created a balanced train dataset that has the same number of sincere and insincere questions. After that, we went through natural language processing: bag of words and we transformed questions from our train dataset into vectors through tokenization. Then we applied the tool named Watson Tone Analyzer to see if a tone of a question can be identified and the identification of tones can have an influence on our prediction. Since a large number of questions canÂťt have certain types of tones(undefined) and also , it was skeptical to use this tool. Through GridsearchCV, we found out that logistic regression have the best accuracy(83%) and cross-validation(84%) and the parameters with C=10 ngram_range = (1,3).

There are some possible directions to extend and elaborate our scope of research. One study direction would be classification of insincere questions into 4 categories which are explained before as 4 main criteria of insincere questions: Non-neutrality of tone, disparaging or inflammatory, not based on reality or a logical assumption, sexual contents. Another suggestion would be classification of answers into sincere or insincere answers. As many as insincere questions exist, there are also toxic and insincere answers. It would be interesting to predict the insincerity of answers.

REFERENCES

[1] Matthew Lynley Frederic Lardinois and John Mannes. 2017. Google is acquiring data science community Kaggle. (2017). Retrieved 2017 from https://techcrunch.com/2017/03/07/google-is-acquiring-data-science-community-kaggle/

[2] IBM. 2019. About. (2019). Retrieved Mar,2019 from https://cloud.ibm.com/docs/services/tone-analyzer?topic=tone-analyzer-about&locale=en

[3] IBM. 2019. The science behind the service. (2019). Retrieved Jun,2019 from https://cloud.ibm.com/docs/services/tone-analyzer?topic=tone-analyzer-ssbts&locale=en

[4] Andreas C. Mueller and Sarah Guido. 2016. *Introduction to Machine Learning with Python.* O'Reilly Media, Inc. http://oreilly.com/catalog/errata.csp?isbn=9781449369415

[5] Quora. 2019. Quora Insincere Questions Classification. Detect toxic content to improve online conversations. (Jan. 2019). Retrieved jul, 2019 from https://www.kaggle.com/c/quora-insincere-questions-classification

[6] Aman Kharb Suman Kalyan Maity and Animesh Mukherjee. 2017. Language Use Matters: Analysis of the Linguistic Structure of Question Texts Can Characterize Answerability in Quora. *ICWSM* (2017). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15647

[7] Jot.S.Singh Sahni Suman K.Maity and Animesh Mukherjee. 2015. Analysis and Prediction of Question Topic Popularity in Community QA Sites: A Case Study of Quora. *ICWSM* (2015). https://doi.org/10.1145/1219092.1219093