Classifiying Insincere Content in Social Media Posts

Kumuda B. G. Murthy
Department of Software Engineering
San Jose State University
California, United States
kumuda.benakanahalliguruprasadamurt @ sjsu.edu

Raghava Devaraje Urs
Department of Software Engineering
San Jose State University
California, United States
raghavadevaraje.urs@sjsu.edu

Shiv Kumar Ganesh
Department of Software Engineering
San Jose State University
California, United States
shivkumar.ganesh@sjsu.edu

***Abstract*— Internet has integrated very deeply in our daily lifestyle. As the internet gets widely available to a variety of audiences, we have seen an explosion in the amount of data produced. Social media specifically has evolved such that day-to-day questions are researched on social media for answers. Each day the data consumption and generation increase exponentially. There are many question and answer forums available for this kind of information sharing, where people from all over the world collaborate to answer the questions. Quora is one such website which provides users the platform to ask and answer questions on any topic. There is no domain restriction on the types of question that can be asked. The challenge these forum faces are the governance of the question posted for any toxic content. The platform deals with questions that carry a non-neutral tone, inflammatory speech, sexual content and questions that are absurd and not grounded in reality. In this paper we propose definitive approach for building machine learning model for systematically classifying the questions as insincere based on the content in the question. The machine learning model will be built using natural language processing on SVM, decision trees and Random forest algorithms.**

Keywords—Quora, internet, Logistic regression, SVM, TF-IDF, Stemming, Decision Tree, Random forest

I. INTRODUCTION

Internet has become the most essential and integral part of current lifestyle. The main aspect for this as how internet makes things accessible and simplifies the tasks that used to be very difficult in past. The day-to-day challenges people encountered are posted and discussed in these social forums for answers or guidance. These questions include wide array of domain like what type of baby product to purchase for a newborn or an immigrant having H1B renewal issues or an employer having legal issues or developer having questions on the problems that he is stuck with and so on. Internet provides a place where anyone in the world can ask a question about anything, which can be answered by anyone in world. Such websites are called as question forums. These forums have gained utmost popularity due to the ease of use. Anyone with basic internet skills can use the forum. Few of the known question forums are Quora, Reddit, Stack Overflow, Yahoo Answers.

The questions posted on social discussion or question forums are answered by experts and novice alike with their personal experience and capacity. Stack overflow is mostly used by the technical community to seek answers which help in their daily coding jobs. Quora, Reddit and Yahoo Answers are similar forums where people ask simple questions which can be personal or professional. These forums provide a platform to collaborate with people with various background and helps in the growth of the community.

Like the two faces of a coin, these platforms too come with a drawback where the users tend to misuse it. Here, the users can post just about anything in a free form text. Some of the questions may be inappropriate, negative sexual aspect, inflammatory, disparaging, hurtful or targeted for a certain set of people. Few of the questions may actually cause mental trauma to the users reading it. There by doing the opposite of what the forums were created for.

The authenticity and credibility of the information is questionable as there are no readily available mechanisms by which websites can perform this validation. Quora is one such platform which will be considered for the purpose of this paper. Quora has several thousand questions each day and would want to classify each question as ‘sincere’ or ‘insincere’. Quora wants to achieve this kind of binary classification to categorize the questions involving[1] non-neutral tone, disparaging or inflammatory, unrealistic fact and use of sexual content on their platform as an ‘insincere question’. To approach this problem, we will be utilizing the [2] data provided by Quora. It has previously tried to do this classification manually to distinguish between the ‘sincere’ and ‘insincere’ questions. We propose a machine learning model that will better serve the purpose rather than manual intervention.

The data provided by Quora is a collection of a training set, test set, and 4 embedding files that are trained on a large corpus. The training data provided has around 1,306,122 unique values whereas the test data has 375,806 unique values.

We will be using supervised machine learning techniques for the above-mentioned problem as the training data is already pre- categorized into ‘sincere’ or ‘insincere’. The machine learning model that we will develop will be a Decision Tree. We will also provide a confusion matrix depicting the outcome of various other algorithms like Logistic regression, Naive Bayes Classification, decision tree, random forest, and SVM. As there is an imbalance in the data, we will be using the F1 score as a measurement metric. This score will provide us a better balance between precision/recall and will be a better measure of the performance of our model.

II. SYSTEM ARCHITECTURE

The proposed system consists of the steps – Data preprocessing, Feature engineering, Modeling, Evaluation and model persistence. We will discuss each of this step-in detail in coming sections. Figure 1 represents the system pictorially.

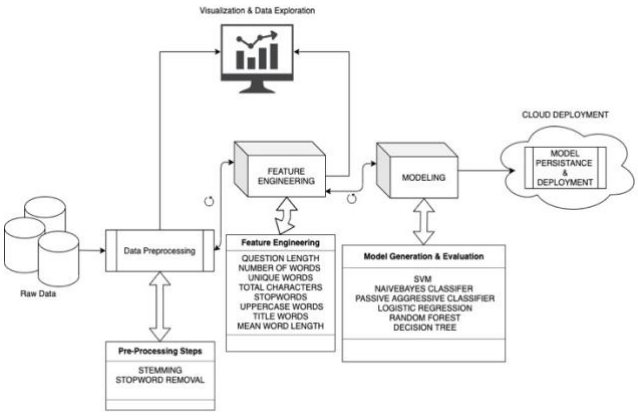

Fig 1. System architecture

III. The Data

The most important task of a machine learning project is to have a good dataset and apply the right methodologies to it. Throughout the paper, we follow the various steps in KDD to classify the Quora questions. For this paper, we collected the Quora data available in Kaggle. Kaggle is a platform which provides many data sets for public use. The data under consideration can be found at the link - https://www.kaggle.com/c/quora-insincere-questions-classification/data .

A. Data exploration

First step is to understand the available data and use it for training the models. The data collected from Kaggle has the following parameters,

- Q_ID – This is the unique identifier assigned to each question.
- Question text – This is the actual content of the question.
- Target – This can be either 0 or 1. Where 1 represents a question that is insincere and 0 represents a question that is sincere.

If a question is sincere or insincere is decided based on the following,

- Is disparaging or inflammatory.
- Is far from reality.
- Has non-neutral tone.
- Has sexual content like pedophilia, incest.

B. Data Preprocessing

- *Tokenization* – Tokenization is a very important step where we split a sentence into multiple words. A package called Natural Language Toolkit is used. This package takes in a sentence and breaks it into individual words. It considers the special characters as separate words. At the end of this step, we will have a list of comma-separated words.
- *Stop word removal* – Stop words are the words used in a sentence to generally join the words and make the sentence more meaningful, like “an”, “in”, “of”. These words so not have any value when considered alone, hence it would be better to remove such stop words. If left untouched, these stop words can influence the results in the modeling step. Once the stop words are removed, the next step would be stemming.
- *Stemming* – There are certain words called root word using which other words are formed. For example, learn is a root word for learning, learnt, learns. Using root words in for modeling will give a higher accuracy. Stemming is one of the techniques used to link words to root words. Using stemming we identify the variants of a same base word.

C. Feature Engineering

- *Common words* – In the English language we have a certain number of words that are commonly categorized as insincere words. Questions having these words are generally insincere.

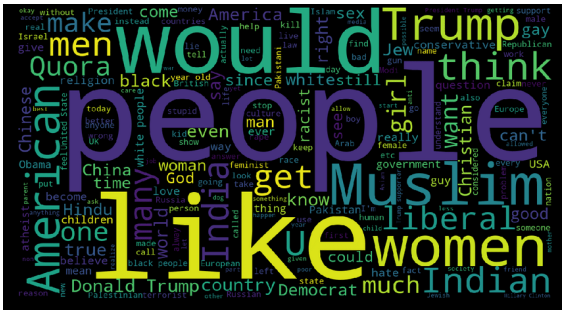

Fig 2. Common word cloud[3]

- *Vectorization* – In this step we convert textual data into numeric data. The most popular approaches for vectorization are TF-IDF and DTM. In this paper we consider Term Frequency Inverse Document Frequency (TFIDF) method.

Term Frequency represents the frequency at which a word occurs in the text. It can be expressed in the following formula[4],

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

Where, j = document, I = term and n represent the frequency.

Inverse Document Frequency represents the importance of a word in degrees. It can be expressed in the following formula[4],

$$IDF_i = log(\frac{N}{n_i + 1} + con)$$

Where, N = number of documents and con is a constant to avoid IDF becoming zero.

Final TF-IDF is represented as follows,

$$TFIDF = TF_{i,j} x IDF_i$$

At the end of this step, we would have converted all text data into numerical vectors.

- *Data Splitting* – The available data is divided in the 80-20 ratio. 80% of the data is used as training set and 20% of the data is used as testing set.

IV. Modeling

Once the data has gone through the preprocessing and engineering phases, we have the training set and test set ready for modeling stage. In this phase we subject the training data into various machine learning models and compare the results.

A. *Naïve Bayes*

This model predicts the result in the form of 0 and 1. 1 represents insincere and 0 represnts sincere i.e if the probability greater than 1, we can conclude the data is insincere. First, the probability is calculated for individual words then using matrix multiplication we calculate the probability for the entire question.DTM is used to get the probability of the entire sentence.

B. *Logistic Regression*

In this model we learn the co efficients during the training phase. This model does not assume anything while training. For this TF-IDF and DTM is used get the probability.

C. *Support Vector Machine (SVM)*

SVM model learns from the training data and forms categories from the learning. Next when the new datat os fed to the model, it allocates previously created categories for the new data. For this project, the model form categories based on the words that occurs again and again. The model finally categorises words into sincere and insincere words.

D. *Decision Tree*

This method is based one binary decisions. At any given time, data is passed to one outcome to see if it fits in, if not passed to the second outcome. For the current project, decision tree will decide sincere and insincere questions based on the content of the question.

E. *Random Forest*

Random forest is a combination of multiple decision trees together. Multiple trees are obtained during the training phase. Then output is obtained using the modes in the decision trees. In the current project, multiple decision trees were produced using various word combinations. Once we have the trees, modes of these word trees are used to get the output.

V. Evaluation and Results

Model	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	0.94	0.79	0.53	0.53
Logistic Regression	0.95	0.82	0.68	0.72
Decision Tree	0.94	0.72	0.69	0.70
Passive Aggressive Classifier	0.94	0.76	0.73	0.74
Linear SVC	0.95	0.82	0.69	0.74
Random Forest	0.94	0.47	0.50	0.48

A Model can be evaluated based on the following metrics,

A. *Accuracy* -– Accuracy is used to measure how close the output is to a designated value.

B. *Precision* – It is the ratio of outcomes predicted correctly to the total outcomes predicted. This does not depend on the accuracy.

C. *Recall* – This is the ratio of outcomes predicted correctly to the total outcomes. Recall is also called as the sensitivity.

D. *F1 Score* – F1 score is the harmonic mean of precision and recall. When closely related it is almost the mean of precision and recall.

VI. Conclusion

Using supervised machine learning algorithms, we were able to accomplish the classification of insincere questions in Quora. By considering Naïve Bayes and Logistic Regression as base models we achieved an accuracy of around 94%, we also tried modeling Support Vector Machine, Decision Tree and Random Forest which took an immense amount of time yielding similar results. Later perceptron like Passive-Aggressive classifier bore better recall and F-1 score compared to other models.
Neural network can be implemented to get an accuracy at its best. This can be called out as future step for this project.

Acknowledgment

The authors wish to thank Dr. Chandrasekhar Vuppalapati for the continued guidance and support. This project was part of academic project submitted to Department of Software engineering, San Jose State University.

References

[1] https://www.kaggle.com/c/quora-insincere-questions-classification

[2] https://www.kaggle.com/c/quora-insincere-questions-classification/data

[3] Akshay Mungekar, Prateek Nima, Nikita Parab, Sanchit Pereira. Quora "Insincere Questions Classification" in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, July 2019, pp. 587–592.

[4] P. Liu, H. Yu, T. Xu, and C. Lan, "Research on archives text classification based on naive bayes," in *2017 IEEE 2nd In- formation Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Dec 2017, pp. 187–190.

[5] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[6] C. Liu, Y. Sheng, Z. Wei, and Y. Yang, "Research of text classification based on improved tf-idf algorithm," in *2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE)*, Aug 2018, pp. 218–222.

[7] F. CHIROMA, H. LIU, and M. COCEA, "Text classification for suicide related tweets," in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, July 2018, pp. 587–592.

[8] O. Aborisade and M. Anwar, "Classification for authorship of tweets by comparing logistic regression and naive bayes classi- fiers," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, July 2018, pp. 269–276