Predicting the Sincerity of a Question Asked

Nguyen Minh Tuan

Department of Information Technology

Faculty of Information Technology and Digital Innovation

King Mongkuts University of Technology North Bangkok

Bangkok, Thailand

minh.tuan@itd.kmutnb.ac.th

Phayung Meesad

Department of Information Technology Management

Faculty of Information Technology and Digital Innovation

King Mongkuts University of Technology North Bangkok

Bangkok, Thailand

phayung.m@itd.kmutnb.ac.th

Abstract—The growth of applications in both scientific socialism and naturalism causes increasingly difficult to assess whether a question is sincere or not. It is mandatory for many marketing and financial companies. Many utilizations will be reconfigured beyond recognition, especially text and images, while others face potential extinction as a corollary of advances in technology and computer science in particular. Analyzing text and image data will be truly needed for understanding valuable insights. In this paper, we analyzed the Quora dataset obtained from Kaggle.com to filter insincere and spam content. We used different preprocessing algorithms and analysis models providing in PySpark. Besides, we analyzed the manner of users established in writing their posts via the proposed prediction models. Finally, we show the most accurate algorithm of the selected algorithms for classifying questions on Quora. The Gradient Boosted Tree was the best model for questions on Quora with accuracy that is 79.5%. Compared to other methods, the same building in Scikit-learn and machine learning LSTM+GRU, applying models in SpySpark could get the better answer in classifying questions on Quora.

Index Terms—text, Quora, insincere, sincere, algorithms, question

I. INTRODUCTION

In the fourth industrial revolution (or industry 4.0), the strong development of information technology has led to a variety of requirements for many companies [1]. For example, Yahoo is solving problems with questions that violate the guidelines of the forum. With the variety, the challenge for companies is how to find the best corresponding data for their marketing and financial policy. Classifying text and images is very useful in the increasing usage of social networks. For example, image classification of flowers could help biologists to easily discriminate the difference of many followers in the same species for research purposes. Or recognizing blur text on ancient stone could help archaeologists understand the history language which prehistoric man marked as the origin of the script.

With the situation, we would like to classify the questions on Quora to get the information with Machine Learning. In this paper, the input data is from the website: <https://www.kaggle.com/c/quora-insincere-questions-classification/data>. An insincere question is used as a question intended to make a statement rather than looking for helpful answers. Table. I) shows some sentences that can signify if a question is insincere as follows:

TABLE I
EXAMPLE FOR 5 QUESTIONS IN TRAIN DATA

Question Text	Target
18male/Mom is 48 and divorced How to properly pursue my mother sexually?	1
6 months left how can I complete my JEE porion in that time?	0
A devote Catholic woman went and asked my gay friend why he "sleeps with men", should he rape and murder her because her question made him angry?	1
A mass shooting will happen again in the US, the question is, who's next?	1
African American people want to stop racial profiling. But why do I keep catching black people stealing at my store if they want the stereotype to stop?	1

Firstly, questions are considered non-neutrally if some characters mean exaggerated to emphasize a point some people and they are rhetorical and aim to point a statement about some people.

Secondly, questions have a disparaging or inflammatory tone when they usually give four characters that containing discriminatory ideas against a class of people, seeks confirmation of a stereotype. They are sometimes making disparaging attacks or insults against a specific person or group of people. They have questions where based on an outlandish premise about a group of people or disparaging against a character that is not fixable and not measurable.

Thirdly, questions give non-biased information or contain absurd assumptions.

Fourthly, questions are using sexual content for shock value, and suggest violent answers.

There are many algorithms and architectures that could be used to optimize and learn from a large dataset. They could use the technologies: Text Convolutional Neural Network (Text CNN), Gated Recurrent Units (GRU), Long Short Term Memory (LSTM) or combine them to extract and analyze the words but could not find the best way to analyze when the new open source is more and more popular.

In the paper [2], they combine different preprocessing and feature representation methods in addition to using the chi-square methods to remove irrelevant features. They also show extensive results showing appropriate feature representation

and filtering, in addition, to classify which enhances the accuracy of the prediction process. The paper [2] also showed the bag-of-words representation or model for Quora data. Further, they showed that the stemming process achieved approximately similar performance over non-stemming. For example, using TP feature representation, F1:0.865 vs. 0.867, Accuracy: 0.866 vs. 0.869, Precision: 0.849 vs. 0.846, Recall: 0.882 vs. 0.889 with slightly better performance for non-stemming especially in terms of recall. Finally, they show that Logistic Regression gets better performance against other classifiers followed by linear SVC, Decision-Tree, and Random Forest but they also do not show all performance for prediction with new technology and algorithms developing every day.

In this paper, we would like to choose data to compare and obtain the best classification technique of questions on Quora. We utilized Apache Spark module called PySpark module structured Query Language (Park SQL) for structured data processing to remove irrelevant features. We also showed Spark Machine Learning (Spark ML) representation or model for Quora data running on Python. After that, we build some models for Logistic Regression, Gradient-Boosted Trees classifiers, and NaiveBayes Classifiers with transformation and actions to apply the models for the data to get the best model in classifying questions on Quora [3]. Besides that, we also try another model of TensorFlow combining with LSTM+GRU to prove the most useful and optimal for the classification. Specifically, our goals and objectives are the followings:

Firstly, we explore the role of text preprocessing and feature representation in detecting insincere content from online social media.

Secondly, we examine the performance of different supervised machine learning algorithms (Logistic Regression, Gradient-Boosted Trees classifiers, and Naive Bayes Classifier) in detecting insincere contents using diverse data representations.

We use Spark evaluator to transform the data by selecting columns from the datagram and calculate the feature of calculations (F1-score, Accuracy, Precision, and Recall). We try another model applying TensorFlow, and LSTM+GRU in classifying to compare with the model applying PySpark about the constructing model and the results.

II. LITERATURE REVIEWS

Logistic regression is a special method to predict a categorical response. This is a very important algorithm in a general linear model to predict the probability of output. Logistic regression could also be used to predict a binary outcome by using a binomial logistic regression or using to predict a multiclass outcome by using multinomial logistic regression.

Gradient boosted trees are a popular classification and regression method using ensembles of decision trees. Gradient boosted trees repeatedly train decision trees to predict and minimize the loss function. Gradient boosted trees applied for binary classification and regression, especially applied for both continuous and categorical features.

Random forests are ensembles of decision trees. Random

forests combine many decision trees to reduce the risk of overfitting. Random forest is applied for binary and multiclass classification and regression. Random forest is used for continuous and categorical features. Label to predict for random forest could be double type. This character is superior to the other model for predicting.

Decision trees and their ensembles are very useful for machine learning task classification and regression. Decision trees are popular because they are very easy to interpret, handle categorical features, extend to the multiclass classification setting, do not require feature scaling, and can capture nonlinearities and feature interactions. Tree ensemble algorithms such as random forest and boosting are among the top performers for classification and regression tasks. Decision trees are applied for binary and multiclass and for regression, using both continuous and categorical features. The implementation partitions data by rows, allowing distributed training with millions or even billions of instances.

LSTM + GRU: LSTM is a kind of artificial recurrent neural network architecture used in deep learning. LSTM has feedback connections. It could process single data points but also entire sequences of data. GRU is a kind of artificial recurrent neural network architecture used in deep learning. GRU combined information to form a new source annotation vector to generate a context vector. GRU could perform in the task of polyphonic music modeling, speech, and natural language processing similar to LSTM. GRU is shown better performance on certain smaller and less frequent datasets.

Naive Bayes classifiers are a family of simple of probabilistic, multiclass classifiers based on applying the Bayes theorem with strong independence assumptions between every pair of features. Naive Bayes could be trained very efficiently. With a single pass over the training data, it computes the conditional probability distribution of each feature given each label. For prediction, it applies Bayes theorem to compute the conditional probability distribution of each label given an observation.

All these algorithms also support for models to predict outcomes. Finally, we apply class `pyspark.ml.Transformer` to transform data set into another. It is very useful for applying with the new data set and bigger data set after building models to predict [4].

III. DATASET

The dataset used in this paper was obtained from Kaggle consisting of four file descriptions: train, test, submission, and embeddings. There are 2,000 questions for the training set. It was identified as insincere (target=1) or not (target=0). The data set had a total of 2,000 records divided into 80% training set, 20% test set.

IV. RESEARCH METHODOLOGY

In this paper, we approach the new way to explore the role of text preprocessing and feature representation by using the available tools in PySpark. We use Python as a data analytics tool to implement experiments.

We use the class `pyspark.ml.feature.Tokenizer` to convert

the input string as a question to lowercase and split words by white spaces. After that, we apply spark.mllib with the feature vectorization method term frequency-inverse document frequency TF-IDF to reflect the importance of the term to a document in the corpus. In the next step, we apply spark.mllib with classification and regression. This includes sections specific classes of algorithms such as linear methods, trees, and ensembles. Based on the characters of the algorithms after testing, we choose some algorithms which are suitable to classify the questions on Quora such as Gradient-Boosted Trees, LSTM+GRU, Random Forest, Decision Trees, Logistic Regression, and Naive Bayes Classifiers. The first parts we use for the training set (1622 questions) and the second part (378 questions) we use for the test part. All of them and the test.csv consist of four stages (Fig. 1):

- + Tokenization of the questions (Create a list of words of each question) and apply Count Vectorizer to convert the list of words to vectors of token counts.
- + Feature selection (apply term frequency-inverse document frequency).
- + Predict outcome by using models.
- + Using transformation for the testing part and test.csv with model and evaluation.

Fig. 1. Preprocessing data

In the question preprocessing, we evaluate two different preprocessing techniques. we remove stop words by tokenizing the messages, creating a list of words of each message, and applied to stem.

In the feature representations, questions are represented using different features by applying CountVectorizer to convert the list of tokens above to vectors of token counts called rawFeatures. After that, we apply the term frequency-inverse document (TF-IDF) [3] and [4]. The cells of the matrix contain tf-idf values of terms calculated by the formulas (1), (2) and

(3), where is the number of occurrences of the considered term in the document, $f_{t,d}$ is a raw count of a term in a document, $|D|$ is the total number of documents, $|d_j : t_i \in d_j|$ is the number of documents where term t_i appears.

$$t, f_{(i,j)} = \frac{n_{(i,j)}}{\sum_k n_{(k,j)}} \quad (1)$$

$$idf_i = \log \frac{|D|}{|d_j : t_i \in d_j|} \quad (2)$$

$$tfidf_{(i,j)} = tf_{(i,j)} \times idf_i \quad (3)$$

In the classification process, there are two features after preprocessing are rawFeature after vectorizing and features after applied term frequency-inverse document (TF-IDF). After that, we try one more step is collecting rawFeature and Features to new column AllFeatures by VectorAssembler. With this step, we get better results than before.

We choose the feature columns for building model and performance valuation. In this paper, we applied four models like the following: Logistic Regression, Gradient-Boosted Trees classifiers, NaiveBayes Classifier with transformation, and LSTM+GRU to apply the models for the data and we get the first results.

V. EXPERIMENTAL RESULTS

After applying four models, we get the results for Logistic Regression, Gradient-Boosted Trees classifier, NaiveBayes Classifier and LSTM+GRU (Table. II). We see that using Gradient-Boosted Trees classifier has the best results in classifying the questions on Quora. In this data set, we use [2] as check prediction using the F1-measure, accuracy, precision, and recall giving as in (4) and (5) :

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

where TP is positive samples, FP is mistake samples positive, TN is negative samples, FN is mistake samples negative. The precision shows how selective the system is, and the Recall shows how thorough it is in selecting useful items another evaluation metric which is called, F-measure is defined to find an optimal trade-off between Precision and Recall values. This metric is achieved by combining the Recall and Precision values [2] within a single measure to devote an equal weight to each of them (6) and (7) :

$$F_1 - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

After running the models, we also get the accuracy for evaluate shown in Table. 1. In the Table. 1, we could see that the accuracy of model using Gradient-Boosted Trees Classifier

TABLE II
RESULTS IN PYSPARK FOR THE MODELS

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.775	0.865	0.660	0.749
Gradient-Boosted Trees	0.795	0.841	0.737	0.786
Random Forest	0.712	0.876	0.510	0.645
Decision Tree	0.720	0.886	0.521	0.656
NaiveBayes	0.761	0.752	0.799	0.775
LSTM +GRU	0.720	0.780	0.720	0.730

Fig. 2. Diagram of different RNN sequence types

Fig. 3. Recurrent Neural Network model

Fig. 4. Diagram of different RNN sequence types

Fig. 5. Recurrent Neural Network model

Fig. 6. Diagram of different RNN sequence types

Fig. 7. Recurrent Neural Network model

is the greatest to predict the questions on Quora. We also try again with the LinearSVC, RandomForest, and DecisionTree, we also get the answer with lower accuracy. Moreover, we try again the steps with removing stopwords and stemming and we also get lower accuracy. This shows that PySpark analyze the raw data better than the other techniques.

After applying the test partition and target test.csv, we also, get some good results. In the Table. II, we have the prediction for the test set with after building model with train and test. The prediction is 0 show that the questions on Quora is insincere and 1 is sincere. With this model, we could predict for all questions on Quora and classify questions into categories of insincere and sincere. To express the groups

of questions in plotting, we create a vertical bar chart for the target test.csv to show the prediction.

The figures Fig. 2 through 7 show the numbers of prediction, number 0 is corresponding with label 1 and prediction 1; number 1 is corresponding with the label 0 and prediction 1; number 2 is corresponding with the label 1 and prediction 0; number 3 is corresponding with the label 0 and prediction is 0.

We the same method when using the same models in Scikit-learn library established in Kaggle task, we get smaller accuracy values than the results shown in Table. III.

TABLE III
ACCURACY RESULTS IN SCIKIT-LEARN FOR THE MODELS

LogisticRegression	GradientBoost	RandomForest	Decision	NaiveBayes
0.755	0.760	0.735	0.710	0.723

VI. CONCLUSION AND FUTURE WORK

Quora is one of the most popular community Q&A sites of recent times. However, many question posts on this Q&A site often do not get answered [6]. In this paper, we have a new approach to analyzing text data using PySpark with Python language. Comparing the other models, the model in PySpark is more natural and simple and requires less computational time than the model in LSTM+GRU, models in SkitLearn library and we get a better result in classifying questions on Quora [7]. We also test again with bigger data and more models such as Linear Regression, Decision Tree Classifier, Random Forest Classifier and we get the same result that Gradient-Boosted Trees Classifier is the best classifier in classifying question on Quora [9].

With the development of technology, analytics in the text will have a new face to appreciate and adjust the behavior for a company or people in general [5] and [8]. Besides that, analyzing data could help people decide perfectly in choosing a suitable method for curing to get better health. The paper also extracted a small data to analyze (2000 questions on Quora) but we also have a better performance comparing with LSTM + GRU and the other algorithms in discriminant that a question is sincere or not.

REFERENCES

- [1] Ohbyung Kwon, Namyoon Lee, Bongsik Shin, Data quality management, data usage experience and acquisition intention of big data analytics, International Journal of Information managemnet, 2014, page 4-5.
- [2] Mohammed Al-Ramani and Izzat Alsmadi ,Using data analytics to filter insincere post from online social networks, a case study: Quora insincere question, Computer information system falcuty publications, 2020, page 5,7,8.
- [3] Richard A. Plunza, Yijia Zhoua, Maria Isabel Carrasco Vintimillab, Kathleen Mckeownc, Tao Yud, Laura Uguccionia, Maria Paola Sutto , Twitter sentiment in New York City parks as measure of well-being, Science Direct, 2019, page 3.
- [4] Yu T, Hidey C, Rambow O, McKeown K, Leveraging sparse and dense feature combinations for sentiment classification, Science Direct, 2017, page 5,6.

- [5] Maryam Khanian Najafabadi, Mohd Naz'ri Mahrin, Suriyati Chuprat, Haslina Md Sarkan, Improving the accuracy of collaborative filtering recommendations using clustering and association rules mining on implicit data, 2016, page 9-10.
- [6] Maity, Suman Kalyan, Kharb, Aman, Mukherjee, Animesh, Language Use Matters: Analysis of the Linguistic Structure of Question Texts Can Characterize Answerability in Quora, 2017, <http://arxiv.org/abs/1703.04001>.
- [7] Maity, Kharb, Mukherjee, Analyzing the Linguistic Structure of Question Texts to Characterize Answerability in Quora, IEEE Transactions on Computational Social Systems, 2018.
- [8] Roy, Pradeep Kumar, Multilayer Convolutional Neural Network to Filter Low Quality Content from Quora, Neural Processing Letters; Aug2020, Vol. 52 Issue 1, 2020.
- [9] Roy, Pradeep Kumar, Deep Neural Network to Predict Answer Votes on Community Question Answering Sites, Neural Processing Letters; Apr2021, Vol. 53 Issue 2, 2021.