Profiling Social Media Questions

Indranil Pal
ipal2@gmu.edu

Brad Staton
bstaton2@gmu.edu

Yasser Parambathkandy
yparamba@gmu.edu

Joshua Kloc
jkloc@gmu.edu

***Abstract*—Question and answer styled social media forums like Quora get thousands of questions every day. In most of the questions, users are genuinely interested in seeking answers but in few cases, someone may ask questions that are provocative in nature. The wording of these insincere questions is designed to make a statement. Weeding out such questions that make a statement rather than look for helpful answers is a key challenge. The website administrators can make use of an automatic way to flag such questions and maintain a civil and respectful discourse. The major objective of this research paper is to predict whether a question is sincere or insincere using machine learning techniques. A dataset obtained from Kaggle regarding questions posted from users on Quora is used for testing and evaluation. The approach focuses on the implementation of neural networks and supervised learning methods. A new method has been tried in this paper to use sentence embedding to get a contextual vector of a question and feed that to neural networks. The same neural network is also subjected to word embedding input to perform a comparative analysis.**

Keywords—Questions Classification, Sentence embedding, Word embedding, Neural networks, Quora

I. Introduction

A pragmatic problem for any major website today is the need 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.

II. Why is this problem important?

Statista estimates that there are 4.65 billion social media users as of April 2022, roughly 58% of the global population [17]. 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.

III. Literature Survey

The following section will discuss the various works done in this area using similar methodologies.

As pointed out by Nima, Prateek, Nikita Parab, Akshay Mungekar, Sanchit Pereira Mungekar in their research regarding insincere question classification, one of the most important parts in text classification and mining is the preprocessing [1]. The results of the different text classification evaluation indicators show that the TF-IDF algorithm has certain advantages in text classification.

As pointed out by Yoon Kim on paper entitled Convolutional Neural Networks for Sentence Classification has performed a series of experiments with CNN on pretrained word-vector for sentence classification [2]. According to his paper, a little hyper parameters turning and static vectors in a simple model (CNN-statics) perform remarkably well, giving competitive results against the more sophisticated deep learning model that utilizes complex pooling schemes. This CNN model improves upon the state of the art on 4 out of 7 tasks which include sentiment analysis and question classification.

Another such paper proposed a model for classifying tweets[3]. The authors used a Logistic Regression model in order to classify tweets according to the topic. The model first transformed the tweets into vectors, which is similar to the one mentioned about by another paper [4]. 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 [5]. 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 imbalanced [6].

In another research [7], 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 an 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 [8]. 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 [9]. 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.

Abdalraouf Hassana and Ausif Mahmood at the University of Bridgeport have done research on Deep Learning for Sentence Classification. The paper observed that most of the machine learning algorithms require input to be denoted as a fixed-length feature like “bag of words”. They ignore the semantics of word and loss ordering of words. Long Short-Term Memory (LSTM) is used over a pre-trained word vector to capture semantic and syntactic information. In the process of trying to predict whether a question is insincere, they used pre-trained word vector, which was trained on 100 billion words of Google News. The use of a pre-trained word vector offers several advantages. A similar word is clustered together. LSTM is used to avoid the problem of vanishing gradient. In their experiment, they used two datasets for sentiment analysis: Stanford Large Movie Review Dataset IMDB and Stanford Sentiment Treebank (SSTB). The training was done through stochastic gradient descent over shuffled minibatches. The size of the hidden state was to be 128 and the mini-batch size was 64. Dropout was set to 0.5 and 10% of the training data was taken for validation. Their model provides a 14.3% error rate for SSTB and an 11.3% error rate for IMDB [10].

Ashwin Dhakal and his co-authors, in their paper - Exploring Deep Learning in Semantic Question Matching has implemented Artificial Neural Network approach to predict the semantic coincidence between the question pairs, extracting highly dominant features and hence, determining the probability of question being duplicate in Quora. In their research work, the words and phrases are mapped into vectors of real numbers followed by feature engineering, which includes NLTK mathematics, Fuzzywuzzy features, and Word mover distances combined with vector distances [11].

Prudhvi Raj, Dachapally and Srikanth Ramanam presented the paper entitled In-Depth Question Classification Using Convolutional Neural Network. According to their paper typically CNN is used for image classification. CNN for NLP is not used often and is completely intuitive. They used two-tier CNN that classifies questions into their main and subcategories. The architecture consists of one Convolutional layer that learns several filters for given heights (Bi-grams to Pent-grams), after that 2-max-pooling layer that accumulates more information from the convolution layer. All the max pooled layers were merged to form a 2-fully connected layer with node 128 and 64. The data used for training was the question classification dataset by the University of Illinois, Urbana Champaign. While testing their model, it was found that 90.43% main category accuracy and 76.52% subcategory accuracy for the Quora dataset which was manually collected. For TREC 93.4% was the main category accuracy and 87.4% subcategory accuracy [12].

A research case study by Aslam et al. [18] offers a considerable comparison point for our own reach efforts. Here, the authors explore the same Quora data set with the same goal to create classification models to predict the sincerity of a given question. Their approach uses both Machine Learning (ML) and Deep Learning (DL) models. The team utilizes logistic regression, support vector machine (SVM), and a long short-term memory (LSTM) neural network. The ML models each use bag of words, bag of n-grams, and TFIDF approach to classify, where as their LSTM Neural Network (NN) is pretrained using GloVe word embeddings. The team then developed three distinct implementations of LSTM classifiers each with different values for layers, hidden units, and dropout rates. This resulted in three variations of the LSTM NN: LSTM, Bi-LSTM, and Deep LSTM models. The team’s ML models generally performed with F1 scores within the 70% range. The best performance came from the SVM using bag of words at 78%. The weakest performing model, at 46%, was the logistic regression with bag of n-grams. Average F1 scores across all ML models was 67%. The DL models outperformed the ML models in terms of F1 scores. The Deep LSTM model performed the best, with 82.5%. The Bi-LSTM model resulted in the lowest F1 score of the DL models at 78%. DL models performed with an 81% score on average.

IV. Methodology

The common aspect in most of the existing research work is the use of word embeddings to represent text before feeding to machine learning techniques. Our models will use sentence embedding with the goal of better performance than that of a word-based embedded modeling approach. Recent work has demonstrated strong transfer task performance using pre-trained sentence level embeddings compared to word embeddings [15]. In this research, we have attempted to use pre-built sentence encoder models to vectorize questions. A comparative study between the two embeddings when subjected to supervised learning and neural network based learnings has been presented.

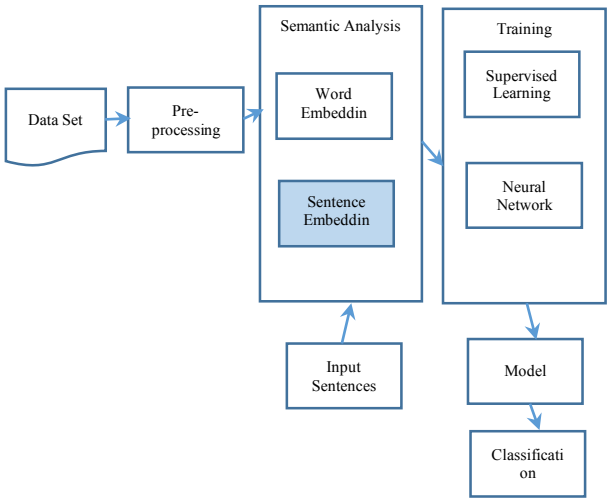

Fig. 1. End-to-end flow of events to clasify questions

"Fig.1" shows the end-to-end flow of our methodology to classify toxic questions. The following sections describe the methodology used to classify toxic questions – Understanding the dataset, data preprocessing, word embedding, sentence embedding, supervised machine learning models – Random Forest, Logistics Regression, and Design of Neural network.

A. Dataset Description

Quora provided dataset contains 1,303,122 questions. The dataset contains three labels qid, question_text, and target which is a binary value and is labeled as 1 for an insincere question or 0 otherwise. Similarly, Quora has also provided out a test dataset that contains 375,806 of test data which contains only two labels qid, question_text. The Kaggle competition allows only word embeddings for the competition and external data is not allowed. Our methodology will use sentence embedding and so we cannot submit classification output to Kaggle site and get an F1 score. For this reason, the training data has been split into train and test data. The training data is further split for validation purposes.

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. Example of sincere and insincere questions are shown in Fig.2

qid	question_text	target
00002165364db923c7e6	How did Quebec nationalists see their province...	0
000032939017120e6e44	Do you have an adopted dog, how would you enco...	0
0000412ca6e4628ce2cf	Why does velocity affect time? Does velocity a...	0
000042bf85aa498cd78e	Why do these idiots keep listening to Steve Harvey for relationship advice?	1
0000455dfa3e01eae3af	How does everyone on Quora seem to be a genius...	1

Fig. 2. First five rows of dataset using pandas

It is worth noting that the classifications of the sincerity, or lack thereof, of questions carries a considerable degree of subjectivity itself. For the purposes of this study, we will take the classification of these questions at face value. Additionally, the dataset does contain noise; the training set data is not completely accurate regarding the classification.

B. Data Analysis

Dataset analysis shows that there are 1,225,312 number of questions that are sincere, labelled as 0 and 80,810 number of the questions are insincere, labelled as 1.

Bar plot of dataset:

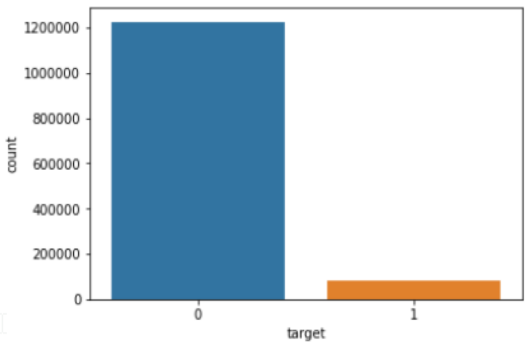

Fig. 3. Number of Sincere (0) and insincere(1) questions

From the bar plot in "Fig.2", it is seen that the dataset contains 93.81% sincere questions and 6.18% insincere questions. There is a class imbalance problem, but we will not attempt to fix it because in Quora website, only few toxic questions appear. The models will be tuned to handle this scenario.

Both sincere and insincere questions have been developed into word clouds, as shown in the figures 4 and 5 below. These word clouds visually convey the text that occurs more frequently

within the data, with more common terms being presented in larger font size.

Anecdotally, there seems to be a more positive tone to words in the questions deemed sincere as evidenced by the prominent use of words 'Best' and 'Good' within the data. The frequent use of those same words does not exist within the insincere data. Not surprisingly, more commonly occurring words within the insincere data have more of an adversarial tone and focus on topics such as ethnicity, nationality, gender, politics, and/or religion. Interestingly, in both cases, the use of the word 'People' is of exceptional focus.

Fig. 4. Wordcloud of sincere questions

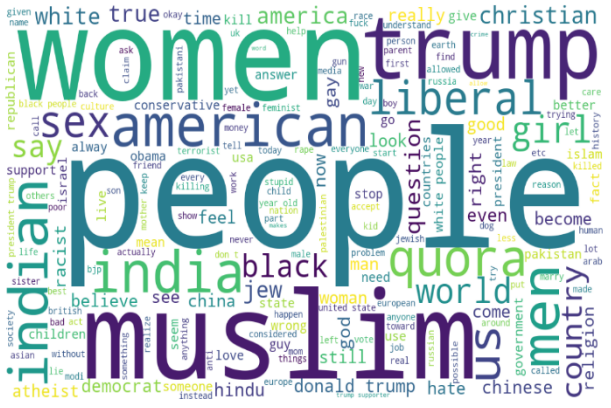

Fig. 5. Wordcloud of insincere questions

C. Data Preprocessing

Preprocessing of data should yield better quality of word embeddings on the dataset. Data preprocessing is a task that includes preparation and transformation of data into a suitable form. Data preprocessing aims to reduce the data size, find the relation between the data, normalize data, remove outliers and extract features for data. The steps necessary to carry out under preprocessing includes - removing punctuations, numbers, stop words like "is", "are", and lemmatization. The goal of lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form [13]. If a question contains "caring" as a word, lemmatization will convert it to "care" and if another question has a word "cared", that is also converted to "care". This might mean the two questions may have similar context.

D. Word 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 [19]. We have used Term Frequency- Inverse Document Frequency (TF-IDF) as the word embedding technique. Term Frequency is the frequency of word in the text. The formula for term frequency is shown below[14]:

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

where, j is the document and i is the term. $n_{i,j}$ is the frequency of repetition of a word in the document. The denominator is the sum of all word's frequency. Inverse Document Frequency (TDF) is the importance of a word denoted in degrees. The formula is shown below:

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

where, N is the number of documents and n_i is the number of documents with the feature word. To prevent the denominator from becoming 0, 1 is added, and con acts as the constant to avoid IDF from becoming 0. Hence, the formula

$$TF\text{-}IDF = TF_{i,j} x\ IDF_i$$

E. Sentence Embedding

An illustration of sentence embedding being better for context awareness for question classification is provided as follows. The question "Is the duck swimming?" and "Should I duck when a cow attacks" should be treated as contextually two different questions. The word duck has different meanings based on the context. Word embedding may treat the two equally, but sentence embedding doesn't.

Google research department created models for encoding sentences into embedding vectors [16]. We use "Universal Sentence Encoder" model for creating embedding of all questions in Quora's dataset. Classification will be performed by finding semantically similar sentences.

F. Supervised Machine Learning Models:

We have implemented Random Forest and Logistic Regression based models. Logistic Regression is a model where the co-efficients are learned during the training of the model. Random forests are a group of decision trees working together for classification of questions. Four models are created.

G. Neural Network Models:

We have used RNN for making model. RNN is a type of neural network in which the output from the previous step is fed as input to the current step.

V. Evaluation and Results

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.

Performance	**Logistic Regression**	
	TF-IDF	***Universal Sentence Encoder***
Precision		
F1-Score		
Recall		

Performance	**Random Forest**	
	TF-IDF	***Universal Sentence Encoder***
Precision		
F1-Score		
Recall		

Performance	**RNN**	
	TF-IDF	***Universal Sentence Encoder***
Precision		
F1-Score		
Recall		

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.

VI. Project Schedule

A. Plan

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

B. Notes

This is not the final research paper. Additional content will be added after implementation of our supervised and deep neural network based models. Along with that implementation we plan to evaluate the results in more comprehensive manner.

References

[1] Nima, Prateek, Nikita Parab, Akshay Mungekar, Sanchit Pereira. (2019). Quora Insincere Questions Classification.

[2] Y. Kim, ―Convolutional neural networks for sentence classification, ‖ arXiv preprint arXiv:1408.5882, 2014

[3] S. T. Indra, L. Wikarsa, and R. Turang, "Using logistic regression method to classify tweets into the selected topics," in 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Oct 2016, pp. 385–390

[4] Fan, H. and Qin, Y., "Research on Text Classification Based on Improved TF-IDF Algorithm" in Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018), May 2018, pp. 501–506

[5] Rafael Geraldeli Rossi, Alneu de Andrade Lopes, Solange Oliveira Rezende, Optimization and label propagation in bipartite heterogeneous networks to improve transductive classification of texts, Information Processing & Management, Volume 52, Issue 2, 2016, pp. 217-257

[6] Aggarwal, C.C., Zhai, C. (2012). A Survey of Text Classification Algorithms. In: Aggarwal, C., Zhai, C. (eds) Mining Text Data. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_6

[7] Yan-Shi Dong and Ke-Song Han, "A comparison of several ensemble methods for text categorization," IEEE International Conference onServices Computing, 2004. (SCC 2004). Proceedings. 2004, 2004, pp. 419-422, doi: 10.1109/SCC.2004.1358033.

[8] A. Wibowo Haryanto, E. Kholid Mawardi and Muljono, "Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification," 2018 International Seminar on Application for Technology of Information and Communication, 2018, pp. 229-233, doi: 10.1109/ISEMANTIC.2018.8549748.

[9] F. CHIROMA, H. LIU and M. COCEA, "Text Classification For Suicide Related Tweets," 2018 International Conference on Machine Learning and Cybernetics (ICMLC), 2018, pp. 587-592, doi: 10.1109/ICMLC.2018.8527039.

[10] A. Hassan and A. Mahmood, ―Deep learning for sentence classification, 2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT). 2017

[11] A. Dhakal, A. Poudel, S. Pandey, S. Gaire, and H. P. Baral, ―Exploring Deep Learning in Semantic Question Matching, presented at the 2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS), Kathmandu, 2018, pp. 86 - 91. doi: 10.1109/CCCS.2018.8586832

[12] Dachapally Prudhvi Raj, — In-depth Question classification using Convolutional Neural Networks, arXiv preprint arXiv:1804.00968, 2018

[13] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008, “Stemming and lemmatization”, pp. 86-91.

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

[15] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data

[16] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

[17] Published by Statista Research Department, & 7, J. (2022, July 7). Internet users in the world 2022. Statista. Retrieved July 10, 2022, from https://www.statista.com/statistics/617136/digital-population-worldwide/

[18] Aslam, I., Zia, M.A., Mumtaz, I., Nawaz, Q., Hashim, M. (2021). Classification of Insincere Questions Using Deep Learning: Quora Dataset Case Study. In: Xu, J., García Márquez, F.P., Ali Hassan, M.H., Duca, G., Hajiyev, A., Altiparmak, F. (eds) Proceedings of the Fifteenth International Conference on Management Science and Engineering Management. ICMSEM 2021. Lecture Notes on Data Engineering and Communications Technologies, vol 78. Springer, Cham. https://doi.org/10.1007/978-3-030-79203-9_12

[19] Brownlee, Jason. “What Are Word Embeddings for Text?” Machine Learning Mastery, 7 Aug. 2019, https://machinelearningmastery.com/what-are-word-embeddings/.