Additional Lit Review / Related Work

A research case study by Aslam et al. 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.

The word embedding is a key difference between this team’s research and our efforts. Our models will use sentence embedding with the goal of better performance than that of a word-based embedded modeling approach.

Data Review / Data Exploration

The dataset being used for this analysis has been retrieved from Kaggle and is associated with an open-source competition to classify questions posted to Quora. The entire dataset weighs in at more than 6.5GB in size. The training dataset contains more than 1.3 million records relating to the questions posted on Quora, each having 3 features. The data is highly imbalanced as nearly 94% of the training data is considered a valid, sincere question with the remaining 6% categorized as insincere. **(Do we want to make any comments here regarding what we plan to do to address the imbalanced data? [Are certain models more susceptible to imbalanced data than others? Is it/is it not a concern in our case?])** Table **X** below breaks out the training data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Label** | **Meaning** | **Count** | **Proportion** |
| 0 | Sincere Questions | 1,225,312 | 93.8% |
| 1 | Insincere Questions | 80,810 | 6.2% |
| Totals |  | 1,306,122 | 100.0% |

Table **X**

Selected examples of both sincere and insincere questions:

Sincere:

* Why is India actually called as India?
* Why do the US Marines have bigger infantry squads than the US Army?
* What’s your favorite kind of cat?
* What little known or small scale war changed the tide of history?
* What is the best email auto-reply software bot?

Insincere:

* How do I lobotomize myself?
* How do I marry an American woman for a Green Card? How much do they charge?
* Why do these idiots keep listening to Steve Harvey for relationship advice?
* What do Americans who think America is the only free country in the world think Europe is like?
* When will Pakistan return Pakistan back to India?

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.

Both sincere and insincere questions have been developed into wordclouds, as shown in the figures below. These wordclouds 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.

Sincere wordcloud:

Text

Description automatically generated

Figure **X**

Insincere wordcloud:

A close-up of a dollar bill

Description automatically generated with medium confidence

Figure **X**