- Random Forest Classifier and SVC take a lot of time to train. Both of them require smaller datasets/a lot of time

Quora Insincere Questions Classification

1. Feature engineering
    a. H Word2Vec
    b. Embeddings, similarity measures
    c. TFIDF N-grams (1-8)
    d. Typographical features (character count, # of capital letters, question length, # of question marks)
    e. Other (time of submission,
    ~~f. Readability (not a good feature)~~
        ~~i. Readability (flesch-kincaid, gunning fog, automated readability index, coleman-liau index, linsear write formula, dale-chall redability score)~~
2. AIs:
    a. Kevin will work on readability-related features
    b. Chris will do the Word2Vec, invite ckt624
    c. Chetan Will work with Typographical features
    d. Soham will look at the TFIDF N-grams
    e. Kongtao will work at similarity measures
    f. Need someone to make a PSDG repository. Once done, we can put the code for the feature engineering there.
        i. GitHub user names: Chris Stewart (cmstewart), Soham Parikh (sohamparikh94), kongtao(ckt624), Chetan(chetan-tutika)
        ii. Gmail accounts: Chris Stewart (stewart.christophermichael@gmail.com), Kevin Sun (kevinsun0@gmail.com), Soham Parikh (sohamp@seas.upenn.edu), Kongtao (ckt624@gmail.com), Chetan(tchetan@seas.upenn.edu)
3. Ckt624
4. Kongtao

Data Stats:

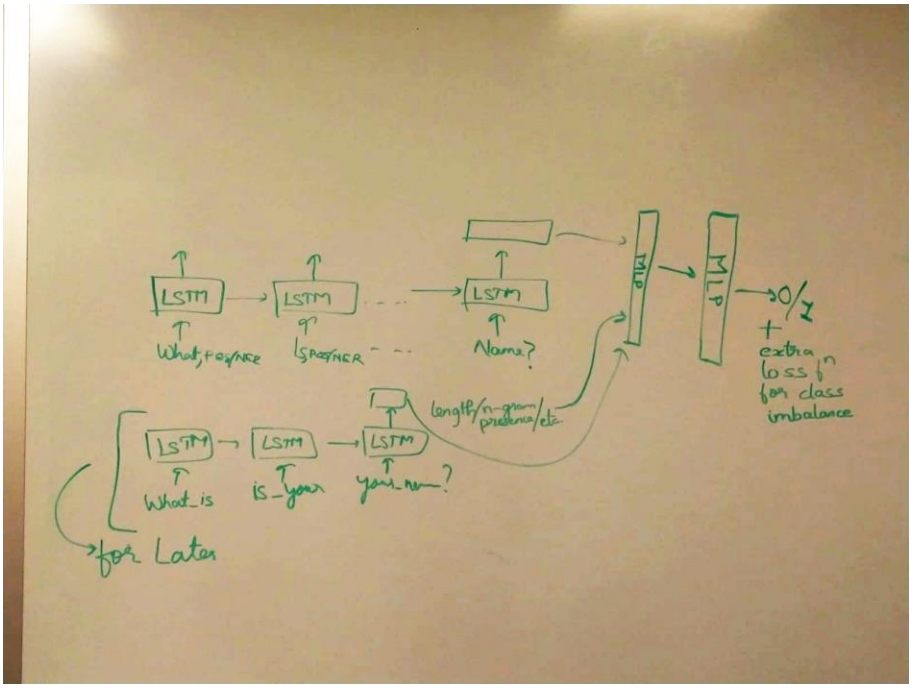
- Train size: 1306122
- Test size: 56370
- Only 6% of the training questions are insincere

Generated Features:

- Word count, No. of capital letters, No. of lower case letters, No. of Question Marks, Stop words in each document, No. of stop words in each document

Week 2:

- Need to split the dataset into test and valid -- Soham
- Perform cross-validation (takes a lot of time so we will leave it for later)
- What embeddings work better? Wiki or Google news may be good since a lot of Quora questions are on current affairs/politics. Paragram vectors can also work better since frequent n-grams tell us a lot about the sincerity of questions. GloVe vectors are the standard embeddings that are used for all tasks.
- Find strategies for class imbalance -
    - SMOTE - can use autoencoder to learn features, followed by creating new class according to SMOTE
    - Cost based function for insincere questions (Kevin)
    - Train it on a balanced set
    - AUC as an objective function when choosing hyperparameters - To be done while training the model
- Code the model: LSTM (word+character embeddings as input at every time step) followed by a FeedForward Neural Network (give additional features also as input) which does binary classification --Kongtao
- Feature Engineering:
    - Find common N-grams for each class (N=1,2,3) (Doesn't make sense to go beyond n=3) -- Kevin
        - Take top 10 n-grams from both classes and check if the question contains that particular n-gram
        - 1-gram (remove stopwords), 2-gram, 3-gram
    - Encode question types as a one-hot encoder -- chetan
        - Who
        - Where
        - How
        - Why
        - What
        - Which
        - When
    - Encode question N-grams -- chetan
        - {How/Why/Who/etc} + {is/did/etc.}
    - Character level embeddings --Soham
    - POS-NER tags --Soham
    - Length of question --Soham
    - Remove stop words from unigrams, but keep them in n-grams for n > 1
    - Number of stop words
- **Model Architecture:**



- [Automatic Identification of Rhetorical Questions]

Week 3

- https://github.com/ksun0/kaggle-quora-insincere
- Collect all features, run feature importance