Project Situation

$$
\begin{aligned}
& OLD\_TEST\_SET + LOG\_LOSS \\
\textit{to}\quad & NEW\_TEST\_SET + LOG\_LOSS \\
\textit{to}\quad & NEW\_TEST\_SET + AUC\_score \\
& \textit{reference}: \ \textit{Kaggle kernels and discussions}
\end{aligned}
$$

Text Cleaning

1. fill NA
2. cut words (nltk TweetTokenizer) ("say." to "say ." but "!!!!!!" to "!!!")
3. convert (i'm, he's ...) to (i am, he is...) (a list online)
4. remove IP address, username, http links
5. remove irrelevant symbols (for now = " $\sim \backslash n$)
6. lemmatize verb (am was to be) and noun (cats to cat, but will convert as to a) using nltk package
7. delete - ("non-degenerate" to "non degenerate")
8. (have not done) correct spelling (kidddddding, pleeeeeese, mothjer etc.) (textblob, but can make mistakes)

Models

1. using word embedding:
   1.1 LSTM (RNN, can also try GRU)
   1.2 CNN
2. using bag of words:
   2.1 (NB)LOGREG
   2.2 (NB)NN

Word Embedding

1. keras text_to_sequence (convert to bag of word then change the text sequence to index sequence) take a max_feature param
2. keras Embedding + GloVe (change index sequence to a list of vectors using GloVe: Global Vectors for Word Representation) + Attention (a dense layer before output)

bag of words

1. tf-idf for words (now use ngram=(1,2)) and characters (now use ngram=(1,5)) with sklearn
2. take nltk english stopping words as stop words
3. words: use top 20,000 and char: use top 35,000
4. next feature engineering

Feature Engineering (Behavior not good)

'word_count', 'cleaned_word_count' , 'unique_word_count',
'cleaned_unique_word_count', 'question_marks' ,
'consecutive_question_marks', 'exclamation_marks' ,
'consecutive_exclamation_marks', 'uppercase_letters', 'ellipsis',
'period' , 'parentheses_pair', 'special_symbol', 'sentence',
'upper_word_ratio' , 'unique_word_ratio', 'mark_count_ratio'

Ensemble

1. using catboost (for now only tried all results as input (bad behavior))
2. plain ensemble (take mean of each column of the results)

Future

1. for text cleaning: find more pattern, remove useless symbols
2. feature engineering: for different label using different features (behave relatively bad on certain labels)
3. column-wised catboost ensemble and add features combined with predicted results
4. grid search for the best params for plain ensemble
5. column-wised CNN, LSTM
6. For new value function, have not tried Naive Bayes to be weight for logistic regression and neural network
7. may need more accounts for submission, and cluster for running deep LSTM, CNN and NN

APPO

"aren't" : "are not", "can't" : "cannot", "couldn't" : "could not", "didn't" : "did not", "doesn't" : "does not", "don't" : "do not", "hadn't" : "had not", "hasn't" : "has not", "haven't" : "have not", "he'd" : "he would", "he'll" : "he will", "he's" : "he is", "i'd" : "i would", "i'd" : "i had", "i'll" : "i will", "i'm" : "i am", "im" : "i am", "isn't" : "is not", "it's" : "it is", "it'll":"it will", "i've" : "i have", "ive" : "i have", "let's" : "let us", "mightn't" : "might not", "mustn't" : "must not", "shan't" : "shall not", "she'd" : "she would", "she'll" : "she will", "she's" : "she is", "shouldn't" : "should not", "that's" : "that is", "there's" : "there is", "they'd" : "they would", "they'll" : "they will", "they're" : "they are", "they've" : "they have", "we'd" : "we would", "we're" : "we are", "weren't" : "were not", "we've" : "we have", "what'll" : "what will", "what're" : "what are", "what's" : "what is", "what've" : "what have", "where's" : "where is", "who'd" : "who would", "who'll" : "who will", "who're" : "who are", "who's" : "who is", "who've" : "who have"...

Current

By, now, our best: 0.9800 (Rank 529) (not with the best models), current LB: 0.9874