Naveen Sukumar
Haoyun Hong
Apoorv Lawange
4/22/22

Reddit Post Classification Model Report

Purpose of NLP Task and Model:

On reddit, there is a subreddit called wallstreetbets where members talk about the status of the stock market and economy and what they are going to invest in. Many people come to the subreddit to learn about the stock market and economy. On the other hand, the subreddit is also filled with memes and other miscellaneous information. We wanted to find out if it is possible to build a recommendation system that can separate informative posts from uninformative posts. As a result we built a model that is able to classify a reddit post as informative or uninformative.

Reporting Accuracy and Confidence Intervals:

Test accuracy for improved model: 0.835, 95% Confidence Intervals: [0.784 0.886]

Test accuracy for baseline model: 0.785, 95% CIs: [0.728 0.842]

Our improved model, which has an accuracy around 0.835, performs better than the baseline logistic regression model, which has an accuracy around 0.785. This shows that the features we implemented in our improved model had a positive effect on the accuracy and performance of the model. Later on in the report is an in-depth analysis of the precision, recall, and f1-score of the improved model we built since that is a better evaluation of a binary classification model.

Explaining Baseline Model vs Improved Model:

The baseline model's features mainly come from the bag of words implementation. Our improved model builds upon the baseline model my adding more robust features under the names of 'econ_term_avg', 'presence_of_curse_words', "presence_of_links_pic", 'links', 'length_of_post', 'curse_words_avg', 'econ_term_counts', 'stats', 'years', 'stock_name', 'org'. Below are explanations of each feature.

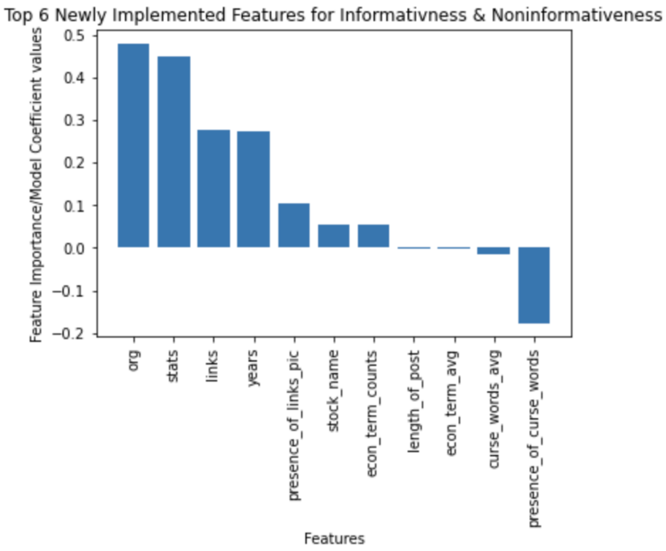
- "econ_term_avg": number of words in post that are econ terminology / total number of words in the post
- "presence_of_curse_words": 0 if the post did not contain a single curse word, 1 otherwise
- "presence_of_links_pic": 1 if the post contained an image, 0 otherwise

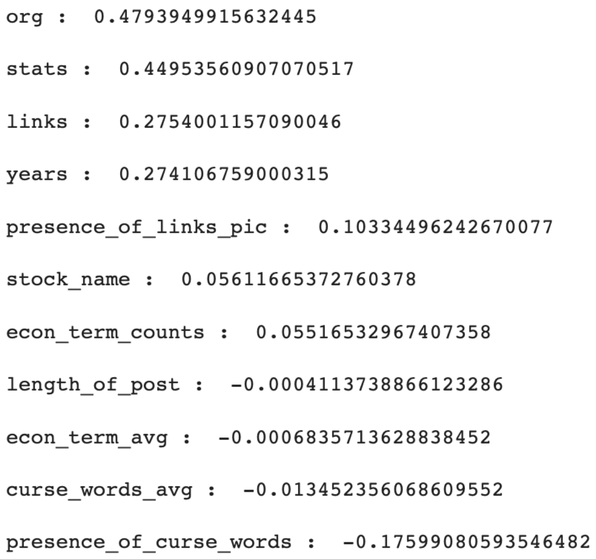
- "links": 1 if the post contained a link, 0 otherwise
- "length_of_post" how many characters long the post is
- "curse_words_avg": number of words in post that are curse words / total number of words in posts
- "econ_term_counts": number of words in post that are econ terms
- "stats": number of words in a post that are highly likely to be associated with facts (such as "increase", "decrease", "percentage", and etc)
- "years": the number of times a year is referenced in the post
- "stock_name": 1 if the post contains a name of a stock, 0 otherwise
- "org": 1 if the post contains the name of an important/powerful entity, whether that is the name of a company, platform, financial institution, or the name of important people.

Background information on both models: Both models output either a 0 or 1. 0 stands for uninformative while 1 stands for informative.

Analyzing Improved Model Feature Importance:

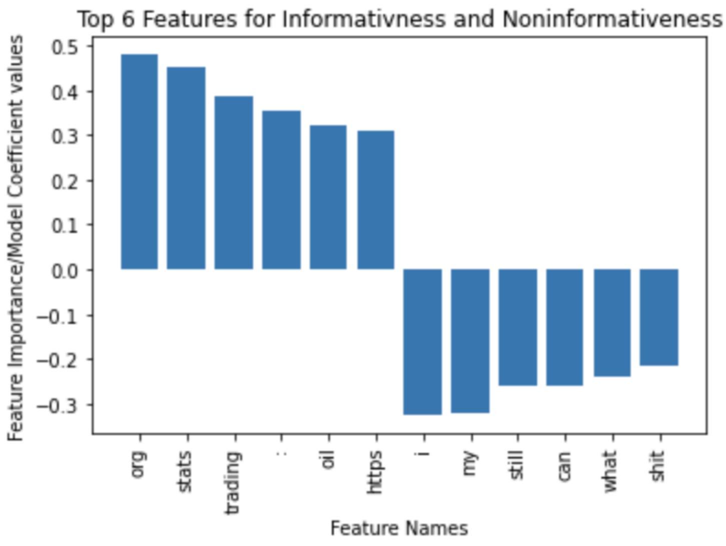
Below are the feature importance values (model coefficient values) of all features we ourselves implemented, which include all features other than the BOW features. Also below is a graph of the feature importance values and a list of the feature importance values.

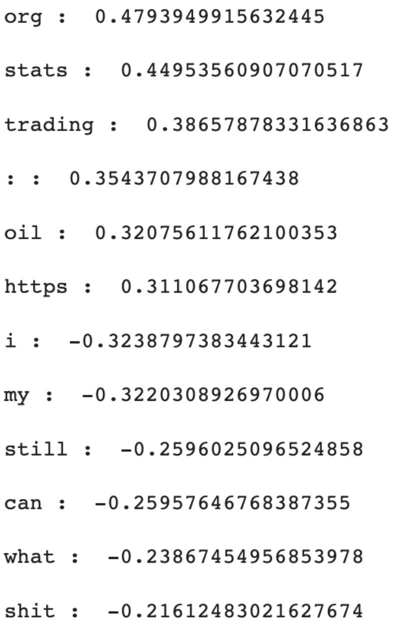



```
org :  0.4793949915632445

stats :  0.44953560907070517

links :  0.2754001157090046

years :  0.274106759000315

presence_of_links_pic :  0.10334496242670077

stock_name :  0.05611665372760378

econ_term_counts :  0.05516532967407358

length_of_post :  -0.0004113738866123286

econ_term_avg :  -0.0006835713628838452

curse_words_avg :  -0.013452356068609552

presence_of_curse_words :  -0.17599080593546482
```

Since the baseline model was a Logistic Regression BOW model while the improved model we developed was the baseline model with extra features that we implemented through feature engineering, we decided to analyze how well our implemented features did. The graph above shows the model coefficients of these features. It can be seen that the “org” and “stats” features performed considerably better than other features. This logically makes sense since informative posts about the economy or stock market tend to bring up the names of companies as well as facts/numbers about their performances. Likewise the feature we implemented that performed well in identifying posts to be uninformative was the “presence_of_curse_words” features. This makes sense because posts that use more curse words tend to be posts that insult others or rant/ramble on. On the other hand, it can be seen that features such as “length_of_post” and “econ_term_present_avg” seemed to have feature importance values very close to 0. We decided to initially featurize the length of a post because posts that are packed with information tend to easily be identified as informative posts. We decided to also featurize the percentage of words in a post that are economic terms because informative reddit posts about the economy/stock market should intuitively use more economic/financial terms. After analyzing the mistakes the model made on the test data, we determined that the "length_of_post" feature was not effective because shorter posts can be packed with useful information and longer posts can be packed with fluff and unrelated information. Moreover, we determined that our “econ_term_present_avg” feature performed poorly because many posts exist that use slang or basic language to explain important information about the stock market or economy.

Now below is a graph and a list of the top 6 features in our model that contributed towards informativeness, and the top 6 features that contributed towards uninformativeness. These include features from the BOW.





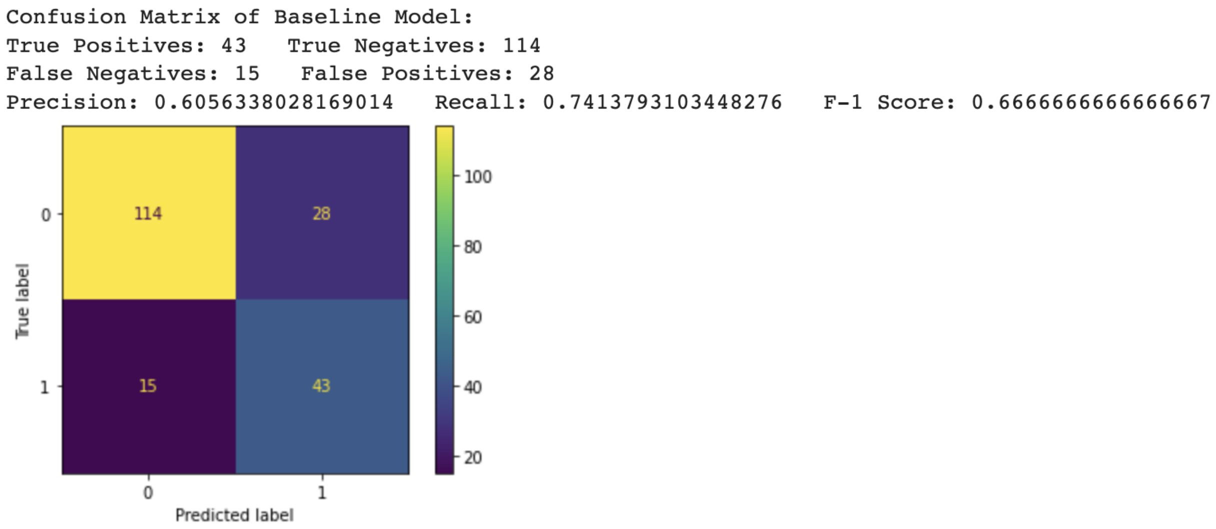
```
org :  0.4793949915632445

stats :  0.44953560907070517

trading :  0.38657878331636863

: :  0.3543707988167438

oil :  0.32075611762100353

https :  0.311067703698142

i :  -0.3238797383443121

my :  -0.3220308926970006

still :  -0.2596025096524858

can :  -0.25957646768387355

what :  -0.23867454956853978

shit :  -0.21612483021627674
```

The graph above shows the top 6 best overall performing features for classifying posts as informative and top 6 best overall performing features for classifying posts as uninformative. For informative posts, it can be seen that the features we implemented, “org” and “stats” had the best performance. Intuitively this makes sense because posts that reference organizations or statistics/data tend to be informative about the economy or stock market. BOW features also seem to have still performed very well with “I” being the best feature for determining if a post is uninformative. This may be because posts that use “I” and "my" a lot tend to be personal stories or rants that are not very informative about the economy or stock market.

An interesting feature to note is that "oil" was part of the 6 top performing features for determining if a post was informative. A logical reason for why "oil" is a higher performing feature is because the oil industry is one of the most influential industries in the economy/stock market and oil is so important that even countries fight over it. This shows that our model has been able to "learn" or identify mentions of oil as signs of the post being highly informative and highly relevant to the economy/stock market.

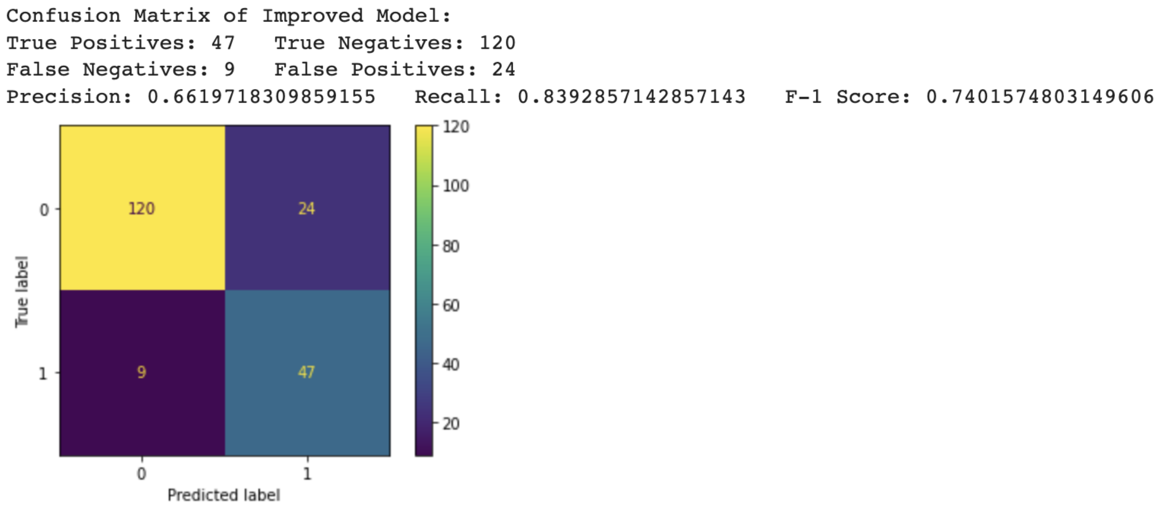
Analyzing TP, TN, FP, FN, Precision, Recall, F1 Score:

Below is the confusion matrix of the baseline model.



```
Confusion Matrix of Baseline Model:
True Positives: 43   True Negatives: 114
False Negatives: 15   False Positives: 28
Precision: 0.6056338028169014   Recall: 0.7413793103448276   F-1 Score: 0.6666666666666667
```

Below is the confusion matrix of the improved model.



```
Confusion Matrix of Improved Model:
True Positives: 47   True Negatives: 120
False Negatives: 9   False Positives: 24
Precision: 0.6619718309859155   Recall: 0.8392857142857143   F-1 Score: 0.7401574803149606
```

Above are two confusion matrices with the quantities of the True Positives, True Negatives, False Negatives, False Positives, Precision, Recall, and F-1 Score printed above.

Taking a look at both matrices, it can be seen from the matrix that there is indeed a class imbalance in our dataset. The test data set contains roughly 129 posts that are labeled as uninformative and 71 posts that are labeled as informative. Hence, there are roughly around 2x more uninformative posts than informative posts. As a result, the improved model has a better chance of finding trends and correlations in the uninformative posts than in the informative posts. Hence, this shows that the results of the model can be biased since its predictive accuracy for the uninformative posts will be lower than its predictive accuracy for the informative posts.

Taking a look at the confusion matrix of the improved model, it can be seen that there are a lot more true positives and true negatives than false positives and false negatives. This is a good sign because it shows that our model has a relatively high accuracy in predicting both labels. Moreover, this shows that our model is able to relatively avoid committing Type 1 and Type 2 errors. The precision of our model is around 0.66, which is above average, while the recall of our model is around 0.83, which is relatively high. Because it is less harmful to classify uninformative posts as informative posts than it is to classify informative posts as uninformative posts, we ideally want the number of false negatives to be lower than the number of false positives, which would make the recall higher than the precision. The recall being relatively higher than the precision shows that our model will tend to produce less false negatives than false positives.

The f1 score is the harmonic mean of the precision and recall of our results. This score is a very useful metric for classification models, especially for situations where there is an uneven balance of labels. The f1 score of our improved model was 0.74, which is relatively high and shows that the model overall performs well for accurately predicting informative and uninformative posts.

Comparing the confusion matrix of the baseline model and the confusion matrix of the improved model, we can see that the improved model is better on all fronts. The improved model has more true positives and true negatives while having less false positives and false negatives than the baseline model. The improved model also has a higher precision, recall, and f1 score than the baseline model. This shows that the new features we implemented caused the model to drastically increase its ability to correctly label the reddit posts.

Analyzing Systematic Mistakes in False Negative Posts

After analyzing these false negative posts, we determined a couple of categories of common mistakes that the model made when classifying these posts.

1. Presence of curse words as signs of insulting or demeaning language

- Example Post: "Alright, retard, all opinions on memestocks and AMC squeeze shenanigans out the window for this post and comment section. We will focus on this AMC gold miner deformed dumpster baby that literally gets sketchier the deeper you really look. But first, how greasy is this deal? Greasy Let us dive in: 1st: Let me take this layup real quick, movie theatre executives don't know dick on how to fucking mine for natural precious metals. Coincidentally, this is about as close as a YOLO a corporate executive/board can go. 2nd: HYMC coincidentally drops an ATM stock offering this morning of up to 500 million unloading on apes as the stocks price launches. For those who don't understand, they are essentially able to soak up this 500% share price increase in taking a shitload of cash while massively diluting their shares. Isn't it odd that this was left out of the AMC press release and not spoken about at all on MSM today? Personally, I read an article in Barrons and NYT that completely failed to even mention it once. Side note: in a total dipshit move that makes Cramer look normal aaron canceled …"
- Many uninformative posts are posts that complain about the economy/stock market or make fun of others for their investments. Such posts tend to use curse words to insult or make fun of others. This post uses a lot of curse words (like "retard") but does not use them to make excessive fun of or insult others. In fact, the post is informative because it brings up a valuable argument that is supported with logic, evidence, and links. Because of the amount of curse words there were in the post, the model seemed to have classified the post as uninformative.

2. Uncommon Acronyms for organizations

- Example Post: 'FB crashing 23% may be an overreaction but it is happing across the market. Pinterest and Spotify crashing at earnings as well. Market is pricing in a bear market for rate hikes. I think the goal should be to slightly average in over the next few months but more so save cash as things may get worse.\n'

- The post above references the company Facebook as FB. Although the model checks for the presence of important companies like Facebook, it is unfamiliar with many of the informal acronyms that people use for such companies. This can cause the "org" feature to not work as well as intended, which can highly likely result in the model misclassifying some informative posts as uninformative.

3. References to events that indirectly affect the economy or stock market

- Example Post: 'I wrote this post last night that started a nice discussion, and thought I would try to keep the conversation going. It seems like Vladimir Putin has decided to reopen negotiations with West. After posturing over 100,000 troops along the Ukrainian border, AP reported that his military has started to pull back some of its forces. So is diplomacy in play? While my analysis looks wrong so far, I did get one aspect correct. Putin was at a standstill and had to make a move, or risk diminishing his bargaining power. I thought Putin would take the opportunity to use aggression and force Zelensky's hand to ultimately reject NATO. He used it to open negotiations on his terms. I recognize that no one wants to go to war. They are expensive and often lead to unintended consequences. But Putin's posturing shift seems to signal something else. If diplomacy is on the table, why not start with that? It could have been a miscalculation, with Putin assuming the Ukrainian comedian would quickly submit. But at the end of the day, Russia wants Ukraine independent from NATO and Ukraine wants to remain in NATO. Both of these realities cannot be true. Putin has to either admit defeat in negotiations, or invade Ukraine. Negotiations will begin quickly, if they occur. We will know if Putin plans to keep his word within the next few days. I'm think we'll hear about a false flag attack in Belarus before that. The markets will react negatively. Positions: SPY puts\n'
- The post references the war between Ukraine and Russia. Even though the post does not contain much economic terms, it references events related to a war that have enormous impacts on the economy or stock market. Because the features of the model highly focus/prioritize using statistics or economic terminology as signs of informativeness, the model mistakenly categorizes posts like this as uninformative.

<u>Analyzing Systematic Mistakes in False Positive Cases</u>

After analyzing these posts we determined 2 categories of common mistakes that the model made when classifying these posts.

1. Randomly mentioning economic entities such as organizations and statistics without explaining their significance to the economy or stock market.

2.Making opinionated statements that can be false or have not been logically supported. Posts contain personal stories or rants rather than relevant information about the economy or stock market

Example Post: 'Let me say up front I am a Contrarian. When people are buying I am slowly selling and vice versa. Been doing it a long time and if you have patience it works. 3 and 4 years ago I was loading up on Energy (VDE, CVX, XON, PSX mainly). And a little 100 year old company that makes wheels and under carriages for big equipment -- TWI. Probably 65% out of then now. What do I think is the next good way to make money with a little patience? https://preview.redd.it/igjx3vfvyii81.png?width=582&format=png&auto=webp&s=69d825d504fa061db35428d30f758fb4013bcafc ARKF, ARKW, and ARK in equal amounts -- buy some every 3 months until tech turns around. Cruise ships also. And hotel ETFs such as INN (get it in the $10 range). We seem to be on the down side of COVID and cruise ships have yet to really take off. Buy all 3, as one may go bankrupt -- not sure which one. And then when these recover, sell and go back into energy and tobacco and wait again. And Best of Luck!\n'

The post above mentions the names of companies but does not talk about their significance to the economy or the stock market. Moreover, the post does not include any logical arguments. The author mentions to take action on stocks with no explanation as for why. Since the model checks for the presence of economic terminology, this could be why the model misclassified these posts as informative.

Future Ideas for Model Improvement

1. Transitioning to using more complex models can be a great way of improving the accuracy/precision of the model. BERT can be a great model to use since it is aware of context and uses "self attention"
2. Using sentiment analysis to determine the intent of a post. Since uninformative posts on reddit tend to very heavily criticize or insult others, we implemented a feature called "curse_word_avg" that computes the percentage of a text containing curse words but discovered that the feature did not perform as well as intended. Using sentiment analysis would be a better way of accomplishing this task.
3. Performing an in depth Exploratory Data Analysis where we take a look at our data and discover trends in the data can be a great way of improving our existing model. The EDA can help us better the feature engineering process and discover new features that can help improve the performance of the model. Conducting a Principal Component Analysis can also help us identify the importance of each feature and determine which features should ultimately be used in the model.
4. Tuning hyperparameters by using K-fold cross validation can improve the performance of the model.
5. Improving the unigram bag of words implementation by using bigrams can make our features more comprehensive and hence improve the model.

6. Using training data over the span of a couple years can greatly improve the performance of the model. Because the training data only consists of reddit posts from this year, the information in the posts could be skewed. Using training data from the span of a couple years can make the training data more diverse and help the model discover new trends. Using data that is balanced between classes might also be an idea to consider.