Classification of COVID-19 Vaccination Side Effects in Tweets

1st Atharva Manjrekar
Computer Science and Engineering
University of Connecticut
Storrs, USA
atharva.manjrekar@uconn.edu

1st Devin J. McConnell
Computer Science and Engineering
University of Connecticut
Storrs, USA
devin.mcconnell@uconn.edu

***Abstract*—Due to the unfiltered nature of discourse within social media platforms, misinformation related to the safety of vaccines continues to rise in our community. This paper analyzes the vaccine dialogue on Twitter during the first 6-12 months of the Covid-19 outbreak. We specifically aim to classify discourse on Twitter surrounding the severity of side effects from the vaccine. Through this framework, we are interested in identifying tweets that show signs of severe side effects with the hopes that we can support such individuals in their time of need. Our classification framework tries to prevent spread of discordant information by distinguishing severe and mild side effects tweets with an accuracy of 68.19% with a Gradient Boosting model carefully crafted with Word2Vec embeddings weighted by TF-IDF. We also show the benefits of state of the art natural language transformer models such as BERT which gives us an accuracy of 80.12% in distinguishing severe and mild side effects tweets.**

I. Introduction

Ever since the Covid-19 vaccines received Emergency Use Authorization (EUA) from the Food and Drug Administration (FDA), we have been undergoing a mass social experiment. The world population is being immunized with a vaccine that was developed in a record setting pace, yet it is not possible to accurately and completely track the range of side effects. The accelerate pace at which this vaccine was developed an rolled out has sparked a need in collecting data on the side effects in various manners. Social media platforms have proven success in crowd sourcing efforts in the past and here it is being used in order to collect the various side effects of the vaccine as they occur. Since most people expect moderate to severe discomfort after going through vaccine shots or treatment plans, they do not believe it is necessary to report any symptoms to medical professionals like the Center of Disease Control (CDC). Many individuals instead turn to social media platforms to post the severity of their side effects and since conversations in such platforms are unfiltered by default, misinformation can rapidly spread around the safety of these vaccines. For example, in a recent poll 42% of respondents cited concern with the Covid-19 vaccine side effects and 12% were unsure about the shot's effectiveness [1].

Therefore, analyzing social media conversations provide insight into the different types of side effects people may have. It also has the potential to reassure people that these symptoms are normal and where they can best get the help they need. In some cases users in these social media platforms may expect certain side effects simply due to the traction a specific severe side effects post may have (likes, retweets, shares, comments) and this can cause vaccine misinformation to spread very rapidly. However, reading about similar experiences people had with very mild to no discomfort could reassure these people on the vaccines effectiveness. Thus, a large number of people may not prioritize getting the vaccine, unless incentivized to do as they may see potential side effects as a major inconvenience to their daily lifestyle/routine. Overall, identifying individuals that have had more severe side effects or have been inconvenienced by the side effects could help public heath officials support the masses and aid educational efforts on reducing vaccine misinformation.

II. Related Work

The Covid-19 pandemic is resulting in a massive infodemic that needs to be addressed. With the large amount of posts being made everyday to spread mis- and dis-information, there is a need to identify various problems within social media platforms. There has been a large effort in annotating tweets and gathering English and Arabic ones into a centralized area [2]. Some of these datasets are large and are multi-lingual consisting of $123M$ tweets [3] and other multilingual datasets are smaller, $1.9K$ tweets focusing on fake information about covid [4]. Others have been focusing on multilingual tweets focusing on tweets with GPS coordinates [5]. Finally, some researchers are looking at the importance of combining news articles and Covid-19 tweets [6].

This infodemic has caused an influx of research on various aspects to Covid-19 and the influence of social media on Covid-19 related topics has been widely studied throughout the pandemic. Amongst the varying work there is an intersection of literature that references vaccine side effects and social media, but limited work has been done in the prediction of Covid-19 vaccine side effect severity through tweets.

A large amount of work has focused on stance detection surrounding the pandemic. We see a popularization of transformer models , i.e. BERT models (mBERT-base, BioBERT, ClinicalBERT and BERTurk), to understand the views of people on vaccination and vaccine types across several countries using Twitter data [7]. Researchers like Cotfas et al. [8] are building

pipelines for Covid-19 vaccine stance detection which aims to help public health entities in all countries provide clear and adequate information on the safety of vaccines by detecting public opinions during the early stages of the pandemic. Under the scope of stance detection we see rumor detection rising in popularity towards Covid-19. Cinelliet et al. is one of the researchers working on rumor amplification over various social media platforms [9]. This work leads into the need of identification of factual information and categorizing it in a variety of labels, False information, Science-based evidence, Fact-checking tweets, Mixed infor-mation, Facts, Other, and Not valid [10]. This pandemic has also caused people to investigate the mis- and dis-information surrounding Covid-19 related to racism and prejudices, labeled as hostile, criticism, prejudice, and neutral [11].

The literature shows that several researchers are working in anti-vaxx and pro-vaxx prediction. Paul et al. [12] investigated opinion mining of Covid-19 tweets for anti-vaxx and pro-vaxx comments in order to classify Twitter data through tree classifiers and neural networks. Researchers have also explored the work of topic modeling, specifically Latent Dirichlet Allocation models, in order to analyze ant-vaxx and pro-vaxx tweets related to Covid-19 [13].

A. Our Contribution

The existing work of Covid-19 and social media lack the fundamental ability to aid in the prediction of vaccine severity. Therefore, we contribute to the prediction of the prediction of Covid-19 severity in several ways:

1) We define a simple method to predict vaccine severity that builds on the pipeline from McConnell et al. [14]
2) Propose interpretable machine learning and neural network methods to predict vaccine severity
3) An analysis of feature importance to vaccine severity tweets

III. Data Collection

Raw data was collected using the Twitter API over the course of a 6-12 month period during the Covid-19 pandemic. This data represents people's reactions and opinions to important events and occurrences during this time period. The data may contain offensive and derogatory language, giving us insight into the emotions that users are feeling. The complete list of features in this dataset and their descriptions can be found in Table: IV.

Examples of the dataset we looked at are seen below, first observing an example of a severe tweet:

I received my second dose of the COVID-19 vaccine yesterday. Boy, the side effects are really kicking my ass. Body aches, small fever amp; chills. But Iâ€™m super grateful that Iâ€™m now fully vaccinated <U+2764><U+FE0F>. https://t.co/xmZSERKY7F

clearly this tweet is discussing the users symptoms of severe body aches and chills. We can also clearly tell the from a second example that someone is experiencing mild symptoms based on their uses of sarcasm, several exclamation marks, and overly jolly mood:

The only side effects we NoKo have experienced so far from the COVID19 vaccine is that i can now see everything in the back of my head which is amazing balls! USA! USA! Thank you @riteaid <f0><U+009F><U+0099><U+008F> https://t.co/BIJHYHNaZm

IV. Data Labeling

All the tweets in our corpus where manually labeled for a variety of features as seen in Table IV. The selected features we chose from the raw data set after calculating feature importance scores are shown in Table I below. Here, the feature that we will predict on is V/L - Severe or Mild side effects. All other features and social metadata was dropped from our final corpus as described in our feature selection section VI-B.

Feature	Description	SF
V/L	Severe or Mild side effects	
Angry	The percent anger seen in the tweets	
Fear	The percent fear seen in the tweets	
Happy	The percent happy seen in the tweets	
Sad	The percent sad seen in the tweets	
Positive	Tweet considered overall positive	
Negative	Tweet considered overall negative	
Neutral	Tweet considered overall neutral	
Favourites Count	Number of tweets a user favorites	✓
Friends Count	The number of people a user is following	✓
Followers Count	The number of users following a single user	✓
Statuses Count	Number of tweets	✓
Listed Count	The number of public lists they are apart of	✓
Favorite Count	Number of favorites the tweet has	✓
Retweet Count	Total number of retweets	✓

TABLE I

Description of Selected Features after feature importance is applied. Feature is the name given within the dataset, Description is what is provided by Twitter API or from our analysis of manually labeled features, and SF is for whether or not it is a social feature.

V. Preliminary Analysis

A. Preprocessing

Any tweets that had missing labels were dropped from our data frame which left us with a total of 1608 tweets in our corpus. Of the 1608 tweets, 59% (949) represent mild tweets and the remaining 41% (659) represent severe tweets.

The textual data in tweets was pre-processed using the NLTK library through Python. The steps we followed are listed as follows:

1) All tweets were converted to UTF-8 encoding and transformed to lower case.
2) The text is then tokenized and all tokens are lemmatized.
3) Stopword removal is used on the lemmatized text.
4) Finally, Normalization is used on the text.

B. Word Clouds

We compiled the words from both severe and mild symptoms categories into word clouds as shown in Fig: 1 and Fig: 2. We analyzed words from within these word clouds, and found associations between them in order to reveal insights into the dialogue of Twitter users pertaining around the severity of vaccine side effects.

We found that both word clouds contained common words such as side effect, vaccine, first dose, covid, etc, which tells us that on average, a tweet from either class will contain these base words. The severe symptoms word cloud contained words that are commonly associated with pain and discomfort such as fatigue, headache, chills, fever, etc. On the other hand, the mild symptoms word cloud shows us very few instances of such words.

C. Social Analysis

Select social parameters were used after conducting feature importance for both severe and mild tweets. Such social metadata gives us an insight to the types of Twitter users that belong to each class. The average values of these parameters are reported in Table II.

Parameter	Severe Tweets	Mild Tweets
Percentage	41% (659)	59% (949)
Favorite Count	13.74	62.09
Retweet Count	0.48	6.16
Followers Count	3524.17	2408.64
Friends Count	1574.2	1756.27
Listed Count	35.11	34.77
Statuses Count	25358.81	27513.93
Favourites Count	31821.43	37599.81
Positive Sentiment	0.44648	0.45935
Neutral Sentiment	0.13150	0.10348
Negative Sentiment	0.42202	0.43717
Happy	0.12882	0.10702
Angry	0.06502	0.08925
Surprise	0.48991	0.50048
Sad	0.16229	0.14728
Fear	0.15306	0.15409

TABLE II
AVERAGE COUNT OF SOCIAL FEATURES PER CLASS.

• **Favorite Count**: This parameter indicates the average number of times a tweet has been liked per class. The average favorite counts for mild tweets is four times more than the average count for severe tweets.

• **Retweet Count**: This parameter provides the average number of times a tweet is retweeted per class. On average, a mild tweet is retweeted about 12.8 times more than a severe tweet.

• **Followers Count**: This represents the average number of followers across the tweet owners per class. On average, twitter users that post severe tweets tend to have more followers than users in the mild class.

• **Friends Count**: This represents the average number of friends across tweet owners per class. Tweet owners in the mild class tend to have slightly more friends than those in the severe class.

• **Listed Count**: The average number of public lists in which users claim membership is about the same for both classes.

• **Statuses Count**: This represents the average number of tweets that are posted by tweet owners per class. Tweet owners for the mild class tend to have more statuses on average than compared to the severe class.

• **Favourites Count**: This parameter indicates the average number of total favorites that twitter users have per class. On average, mild tweet owners are considerably more liked than severe tweet owners.

• **Positive Sentiment**: Average positive sentiment score for both classes. The scores from both classes are roughly the same with mild tweets presenting just slightly more positive sentiment compared to severe tweets.

• **Neutral Sentiment**: Average neutral sentiment score for both classes. Here, we see severe tweets showing a higher neutral sentiment compared to mild tweets. Tweet owners in the severe class also may not know effectively how to react after taking the shot since they may feel the side effects could be temporary and long term effect on their health. So it is understandable to see users in this class have a more conservative sentiment and as such the neutral sentiment is ranked higher here than the mild class.

• **Negative Sentiment**: Average negative sentiment score for both classes. Here, mild tweets show slightly higher negative sentiment compared to severe tweets.

• **Happy**: Average happiness score for tweets in both classes. Severe tweets tend to show more happiness than mild tweets but all round the average scores tend to be lower for both classes.

• **Angry**: Average angry score for tweets in both classes. Mild tweets tend to show more anger compared to severe tweets. Here, anger may be masked as sarcasm as well since jokes and overly hyped reactions can imply both severe and mild side effects. Since it is very difficult to label sarcasm based tweets into either of the classes, by default, such tweets were labeled as mild.

• **Surprise**: Average surprise score for tweets in both classes. Mild tweets tend to showcase a more surprised emotion which may represent their surprise to not receiving any side effects from the shot. The relative high surprise score in both classes show us that users are surprised by the outcome of their vaccine shot.

• **Sad**: Average sadness scores for tweets in both classes. Severe tweets tend to score higher on sadness than mild tweets. This makes sense since severe tweets do tend to express sorrow and resentment when being affected by certain severe side effects from the vaccine.

• **Fear**: Average fear score for tweets in both classes. The close proximity of this score for both classes show the fear that is currently present for tweet owners in both classes. This shows us that no one in both classes truly thinks they may not have long-term side effects and as such are still fearful of the negative effects this vaccine may bring overtime.

On average, tweets from the severe class have more followers than tweets from the mild class. One reason why this

discrepancy may be is due to certain high profile individuals who are heavy vaccine skeptics in general and even the slightest side effect for them could be exaggerated as severe side effect. These individuals tend to get a lot of traction on social media platforms as they are able to spread information on vaccine effectiveness without any sources or credibility to simply put fear and distraught on other people. These tweets designed to instill fear and distraught in people tend to gain traction very quickly. Therefore, it is reasonable to assume most of these people have amassed a huge fan base which on average is greater than the twitter users in the mild class.

Additionally, it's important to understand that tweet owners in both classes are most likely pro-vaccine as they have decided to take the shots before tweeting and someone who is truly anti-vaccine will not have taken their shots at all. This means that while the severe class has tweets with severe severity, these tweets do not necessarily imply anti-vaccine sentiment but could imply genuine severe side effects. However, such uncontrolled tweets can potentially lead to skepticism on the effectiveness of the vaccines and successful classification between the two classes is the first step to mitigating such skepticism.

VI. Feature Engineering

A. Feature Engineering: Vaccine Severity in Tweets

1) Sentimental Features for Tweets: Many works put emphasis on the use of sentimental features like polarity and emotions [15], [16]. The polarity of a tweet is the idea of trying to extract the overall opinion that someone has in regards to what they are saying. Through the use of Vader Sentiment we were able to extract polarity scores, which tweets that score < -0.5 is labeled negative, > 0.5 is labeled positive, and all other tweets are labeled as neutral. We also used text2emotion which will allow you to extract the emotions happy, angry, sad, surprised, and fear from the textual data provided. In this work we keep all of the percentage scores that are provided for each emotion for each tweet.

2) Rule-based features for Tweets: In order to find keywords that are unique to each class (Severe - V and Mild - L), we used a graph based ranking model called TextRank [17]. This model is commonly used for text processing in natural language applications where the results obtained compare favorably with previously published results on established benchmarks. Since TextRank is a graph based model, each word in our class will serve as a vertex within a graph. The units that we will rank are one or more lexical units extracted from processed text. These also represent the vertices on a graph and any relation between two lexical units is represented as an *edge* between two such vertices. Intuitively, TextRank works well because it does not only rely on the local context of a text unit (vertex), but rather it takes into account information recursively drawn from the entire text (graph).

The methodology of our graph based ranking algorithm consists of the following steps:

1) Identify the best textual units and add them as vertices in the graph

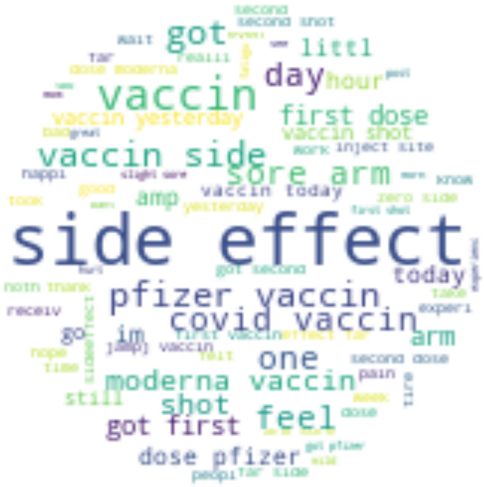

Fig. 1. **Mild Words - Word Cloud, Mild Symptoms Reported in Tweets**

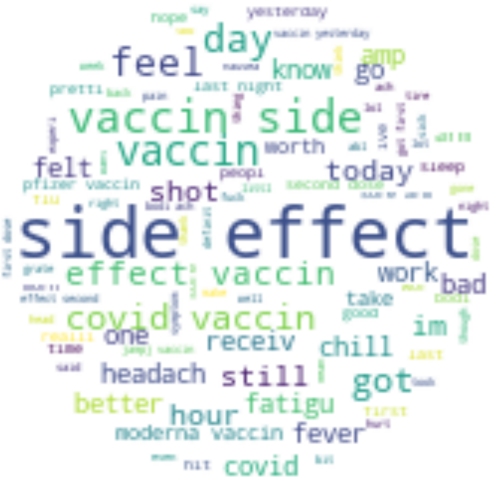

Fig. 2. **Severe Words - Word Cloud, Severe Symptoms Reported in Tweets**

2) Identify relations that connect such text units, and use these relations to draw edges between vertices in the graph
3) Iterate the ranking algorithm until convergence
4) Sort vertices based on their final score

We identified that words scoring above 2.0 from both classes aided more in our classification than those that scored lower, as seen in Fig. 4 and Supp. Fig. 10 there is a spike in the use

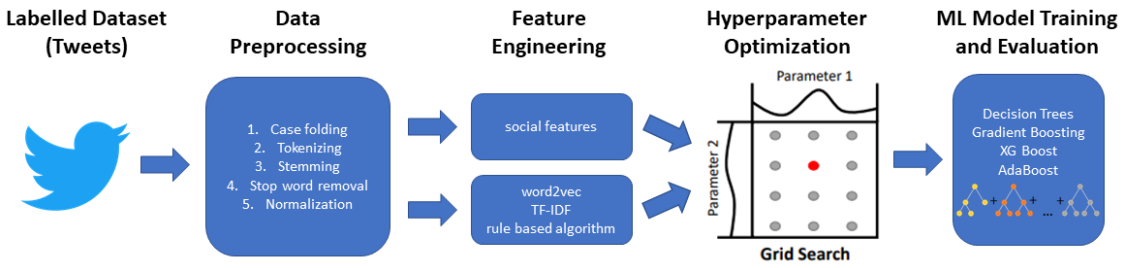

Fig. 3. **Twitter Vaccine Severity Pipeline. Here we follow a similar pipeline as presented by McConnell et al. [14]. The beginning stages of the pipeline we worked with labelled Twitter data for vaccine severity that contains the full set of social features found in Table: IV. We then did data preprocessing on the text data before feature engineering. We went through two sets of feature engineering one on the social features defined in Section: VI and worked on rule learned algorithms, word2vec embeddings, and TF-IDF. Here, the grid search that is done is described in two dimensions where circles denote parameter configurations and curves on the axis denote the classification accuracy. In this example, classification accuracy has higher variability for the values of parameter one, meaning parameter one primarily determines the best setting (red point). This then leads into the four ML models we trained and evaluate for the rest of the paper.**

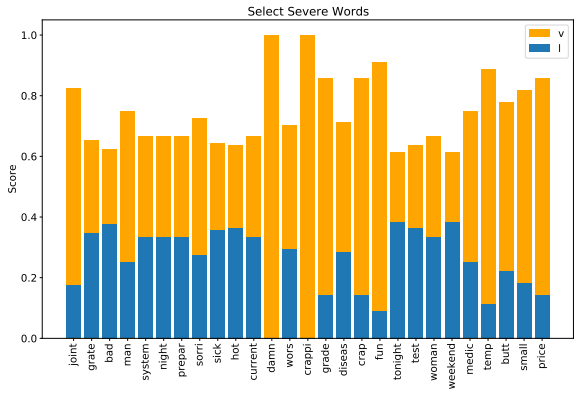

Fig. 4. **Scores For Select Keywords From Severe Tweets. Select keywords from the severe side effects tweets are presented here and scored for there representation in both classes. These keywords are selected from TextRank [17] with a score from there model of greater than or equal to two. The way we scored each word is defined under Sec. VI-A2. The differences between scores show us how diverse the opinions are between the two classes. For example, the severe keywords tend to represent pain, discomfort, and medical terms as shown by words like temp, joint, sick, medi, etc, which have a higher prevalence in severe tweets.**

of words between different classes at this scored value. These values were normalized by counting the number of times the word shows up in one class over the number of times it appears in the entire corpus. This can be seen in an example of severity normalization:

$$\frac{w_v}{w_v + w_l}$$

where w_v is the count of how many times a word, w, appears in class v and w_l is the count of how many times w appears in class l.

3) Word Embeddings for Tweets: In this paper we focus on the the word2vec [18] architecture in order to construct twitter features from the neural embeddings. The objective of the word2vec model is to maximize log $P(w_O|w_I)$, of the log probability of a given word W_O given an input word w_I. For these models we are computing 300 dimensional vectors for each tweet using a word2vec model that was pre-trained on AP News or Wiki articles. We also investigated weighting word2vec models by term frequency inverse document frequency (TF-IDF) as described by Lilleber et al. [19]. The TF-IDF can be defined as:

$$TF - IDF = tf * log\frac{t}{df}$$

where tf is the number of times a term, t, appears in a document and $\frac{t}{df}$ is the inverse document frequency, such that df is how many documents term t appears in. We also produced models that are trained on a combination of word embeddings and rule classifiers [14]. These models compute an average word2vec vector weighted by TF-IDF and includes only those words that are identified by the the rule learner algorithm described above.

B. Feature Selection

Feature selection is a task that requires the subsetting of the initial features when building a ML model. This is done to in order to eliminate the features in the dataset that are not important to the class that you are trying to predict. Here we use an extra tree model and follow the methodology described by Strobl et al. [20] in order to obtain the feature importance and eliminate social features and sentimental features with less than 0.003 importance to the class we are trying to predict. In Fig: 5 we display the feature importance that was generated from the extra tree classifier. Allowing us to reduce the original features seen in Table: IV to the subset features seen in Table: I. This left us with a final set of social features that are most important to the class being predicted.

VII. Modeling Vaccine Severity in Tweets

We considered four classification models to predict tweets discussion on vaccine severity: adaBoost [21], decision trees [22], gradient boosting [23], XGBoost [24], Multilayer Perceptron [25], and Bert with a deep neural network [26].

A. Tree Classifiers

Tree classifiers construct a set of decision rules that split the data until a leaf is reached which signifies the class label.

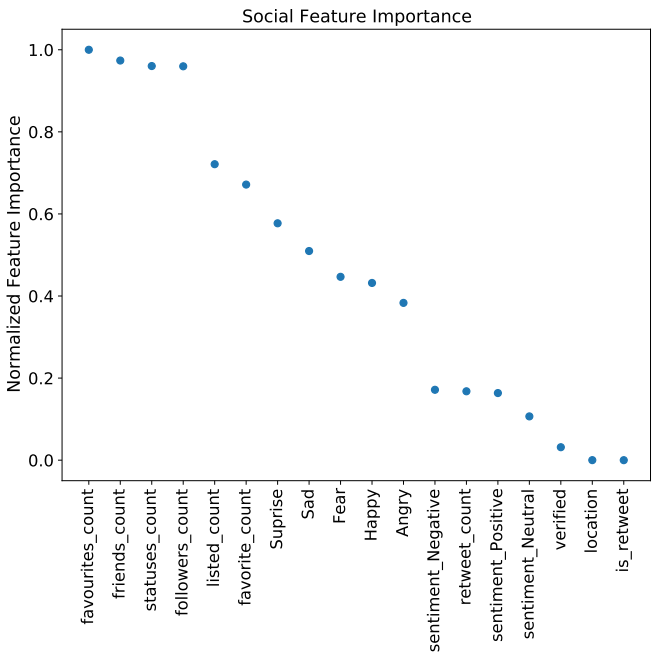

Fig. 5. **Social Feature Importance, Displays the feature importance of the social features of the tweets in relation to an Random Forest Classifier. This model was chosen because of its high interpretability of its features. The vertical axis is the score of the feature and the x-axis gives us the feature. We can see a large spike in importance with favourites count in relation to predicting vaccine severity. The importance progressively decreases to 0.003 at the feature Verified. We use a feature importance of < 0.003 to remove features from our models.**

The hierarchy of rules is rep-resented as a tree where internal nodes denote a bifurcation of a subset of samples that are split based on various of metrics. However, selecting an optimal decision tree is an NP-hard [27], [28] problem causing trees to be constructed in a greedy manor in one of several methods. One method in aiding in selecting these trees is information gain. It is the change between the information entropy at an internal node versus the conditional entropy of a feature. Another method is gini index calculates the probability that a specific feature is classified incorrectly when it is selected randomly.

B. Boosting Methods

Boosting ML techniques are based on building an accurate model from an ensemble of weak models sequentially [29]. Here we train adaboost, gradient boosting trees, and XGBoost classifiers. In adaboost models [30], new decision trees are built with successive boosting iterations re-weighting training instances. This is done so the newly built decision trees focus more on the samples that were previously misclassified. The objective of gradient boosting trees are to approximate the negative gradient of a binomial deviance loss function [23]. Building on the gradient boosting model we used XGBoost, or Extreme Gradient Boosting. It is an efficient implementation of the gradient boosting tree with an adjusted loss function that control's the complexity of the weak leaner (decision trees) [31].

C. Neural Networks

Neural Networks, specifically multilayer perceptrons (MLP), goal is to approximate some function (f). Specifically during the classification model we try to learn a mapping such that $y = f(x; z)$ where x is the input, y is the output and z are the learned parameters to best approximate the function [25]. We also work with Bert [26] whose underlying model for classification is a deep neural network that has randomly initialized layers where each layer between the input and output layers have a different object.

VIII. Performance Metrics

Our main objective is to detect severity of Covid-19 vaccine within tweets, thus, we can define the performance metrics as follows with mild being positive and severe being negative. Tweets can classified into four groups – true positive (TP) (mild labeled as mild), true negative (TN) (severe labeled as severe), false positive (FP) (severe labeled as mild), and a false negative (FN) (mild labeled as severe). These four groups lead to the following metrics to compare classifier performance:

1) *Accuracy* is the percentage of tweets that are labeled correctly. During model selection we prioritized this metric:
$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$
2) *Precision* is the percentage of the tweets that are actually mild severity out of all the tweets that are predicted as mild severity:
$$Precision = \frac{TP}{TP + FP}$$
3) *Recall* is the measures of how many tweets are labeled mild severity and are actually mild severity:
$$Recall = \frac{TP}{TP + FN}$$
4) *F1-score* is the balance between Precision and Recall:
$$F1 - score = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

IX. Results

It has been show that classifier performance greatly depend on model selection [32]. For model selection, we performed a grid search for each method, with 10 fold cross validation on 70% of the data used for training and validation. We report the accuracy of the models from the grid search in Supp. Table V. Then for estimating the variability of the selected models from the grid search we computed 100 bootstrapped samples and reported the median accuracy, precision, recall, and F1-score in Table VI.

To assist interpretation of model performance, we compared our ML models with a naive baseline. The naive classifier predicts vaccine severity within tweets using the empirical

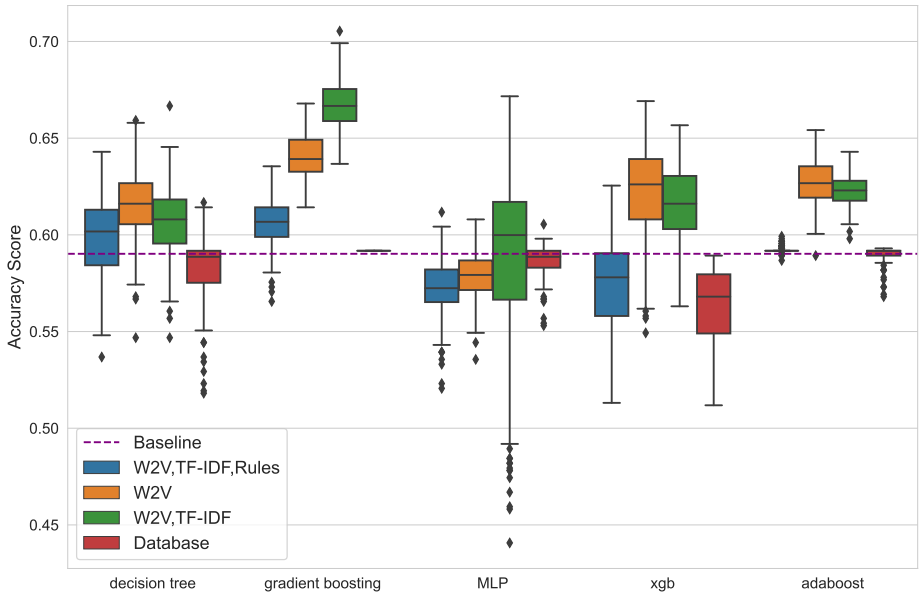

Fig. 6. **Predicting vaccine severity. Here we compare the variance of our top models on different features and compare them to baseline model (purple line). We show the Database feature only models (red), document embedding features based on word2vec [18] (orange), TF-IDF weighted Word2Vec [19] (green), and a rules only TF-IDF weighted Word2Vec model [14] (blue). We see a higher median accuracy when using the Word2Vec,TF-IDF model over the others and in most cases less variability in the models accuracy for the four classifiers: adaBoost (ada), decision trees (dt), gradient boosting (gb), xgboost (xgb), Multilayer Perceptron (MLP). Box plots are drawn with Tukey whiskers (median ± 1.5 times interquartile range).**

frequency of the training set; with about 59% of the tweets being related to mild side effects, we observed a naive baseline accuracy of 0.5902. During model selection, we observed a maximum classification accuracy of 0.6819 using a gradient boosting classifier with dense word embeddings from Wiki weighted by TF-IDF. We also note that the highest median accuracy comes from our Word2Vec weighted by TF-IDF data on a gradient boosting classifier with 0.6667 accuracy. This model also display the highest F1-score.

First, we evaluated and highlight the difference in feature sets developed, Database, W2V, (W2V, TF-IDF), and (W2V, TF-IDF, Rules), in order to determine there ability to predict vaccine severity (Fig. 6). We evaluated all methods, but noticed an increase in accuracy as we built on top of dense natural language features extracted from the tweets. However, we observed a halt in improved accuracy at our model that is crafted by weighting (TF-IDF) word2vec embeddings for our models. This result allows us to focus the rest of out findings on the Word2Vec weighted by TF-IDF models.

We next investigated feature importance for vaccine severity prediction from decision trees built with Word2Vec features. We selected decisions trees since they are the most interpretable model from the four models we ran. Our findings showed that Word2Vec features was universally the most important feature to the models (Fig. 7). This can be interpreted as the textual can accurately capture the meaning of vaccine severity. We also note that the favourtie count feature was the most important feature before dense features were added (Fig. 5) and is the second most important after the addition.

Then we worked on quantifying if severe or mild side effects made prediction easier through stratifying classification accuracy by severity in the best performing model in accordance to the highest median accuracy (W2V, TF-IDF). All methods, predicted mild severity of vaccine side effects with a significantly higher accuracy than those tweets that are labeled as severe side effect (two-way paired t-test, $p \leq 3.9 \times 10^{-15}$) (Figure 8). This is most likely due to the properties of mild side effect tweets and not class imbalance since we do not observe a significant difference in the percentage of tweets labeled as mild, 59%, and those labeled as severe, 41%, in the data.

A final analysis on these models included an investigation on the average precision (AP) in Precision-Recall Curves and the area under the curve (AUC) for Receiver Operating Characteristic Curves. In Fig. 9 we show the AP and AUC from our best performing model comparatively between APNews and Wiki embeddings. We show the AP and AUC for all other models in Supp. Fig. 11 and Supp. Fig. 12. Our highest AP is coming from a gradient boosting classifier with .77 average precision with APNews embeddings and an adaboost classifier on Wiki word embeddings with .77 average precision. The highest AUC for our models are coming from a

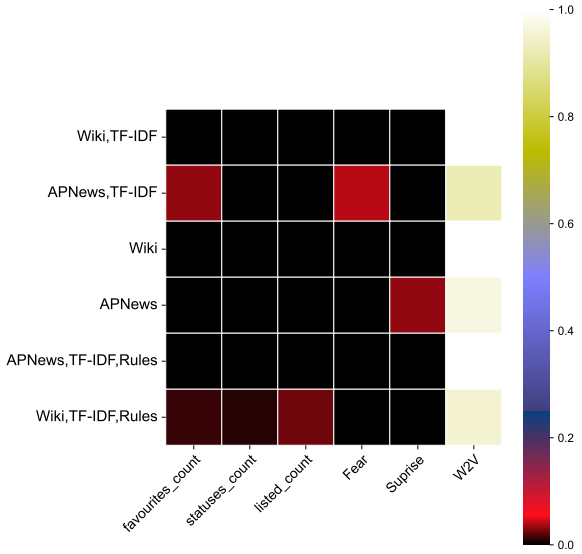

Fig. 7. **Decision tree feature importance for models with word2vec features, the vertical axis denotes the models that were run on the data. Wiki is for Wiki Word2Vec (W2V) models, APNews are APNews W2V models, TF-IDF is weighted W2V, and Rules are models only using the rules from our rule learner. The x-axis gives us the feature importance of all feature that gave us at least one model with a non-zero feature importance. We can see that W2V features have a higher level of importance to the prediction of our class across the board. We also see an increase in importance for the favourites count features in our models. The W2V features are summed across all W2V features.**

gradient boosting classifier on Wiki embeddings with an AUC of 0.7425.

Finally we investigated Bert models in order to explore the effectiveness of natural language transformers. In this model we used a pre-trained Bert transformer that is a case sensitive model and trained on a large corpus that includes Wikipedia articles. We explored different epochs and learning rates on a $70\%/30\%$ train-test split in order to achieve an optimal accuracy. This tuned Bert model produced a higher accuracy, precision, recall, and F1-score as seen in Table: III. Our Bert model was trained independently from the concepts presented in our pipeline shown in Fig. 3 to gather benefits of transformers in text classification.

BERT Model	Score
Accuracy	0.8012
Precision	0.8203
Recall	0.8491
F1 Score	0.8835

TABLE III
SCORING OF PRE-TRAINED BERT MODEL, THIS ANALYSIS SHOWS THAT STATE OF THE ART BERT MODELS PROVE TO BE MORE ACCURATE THAN WORD2VEC MODELS. THIS IS EVIDENT FROM THE ACCURACY OBTAINED WITH A BERT MODEL OF 0.8012 AND OUR BEST PERFORMING MODEL ONLY SCORING 0.6819.

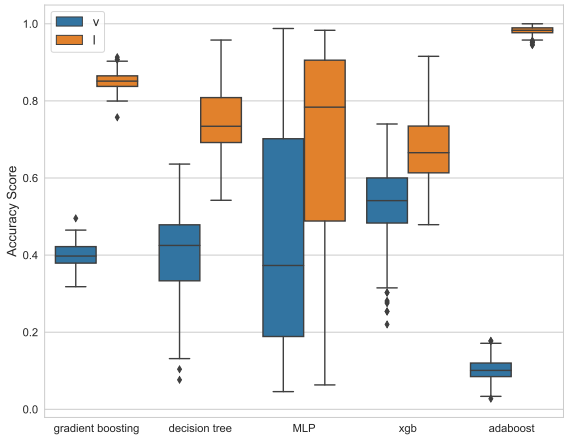

Fig. 8. **Classifier accuracy by severity type. We stratified the classifier performance on vaccine severity type inorder to predict severe side effects or mild side effects. Here, we include our classifiers trained on W2V and TF-IDF features. The box plots show the distribution of 100 bootstrapped samples per Word2Vec embedding model with Tukey whiskers (median ± 1.5 times interquartile range).**

X. DISCUSSION

In this work we focused on predicting vaccine severity, but the prediction of other Covid-19 related work may be of strong interest to the CDC. For example, prediction of emotions, misinformation, and hesitancy. We understand that careful consideration is needed when deciding which features are most effective during feature engineering of social and textual data. For example, we could work with different n-grams, utilize variations on the social features, look at users profile descriptions, and any links to improve scores of our performance metrics.

The comparisons we performed between Word2Vec, Word2Vec weighted by TF-IDF, and Word2Vec weighted by TF-IDF for only rules shows the benefit of features involved in textual features. Our additional exploration of Bert models in Table: III shows that there is an additional benefit in exploring modern transformer architects that work on domain specific pre-trained Bert models, e.g. BioBERT [33]. Models that are domain specific have proven to be more successful than those that are fine-tuned from the general corpus. These types of architectures will be acceptable to attempt a domain specific vaccine severity on tweets, however extending it from twitter to posts on other platforms which may be larger would benefit from the recent work on extending the range of transformers [34], [35].

A limitation in the data is the inability to distinguish between actual mild severity and someone joking about vaccine side effects. With the growing interest of vaccine side effects by the CDC it is important to be able to distinguish between severe and mild side effects with another class that contains sarcasm or jokes on vaccine severity. There is a large

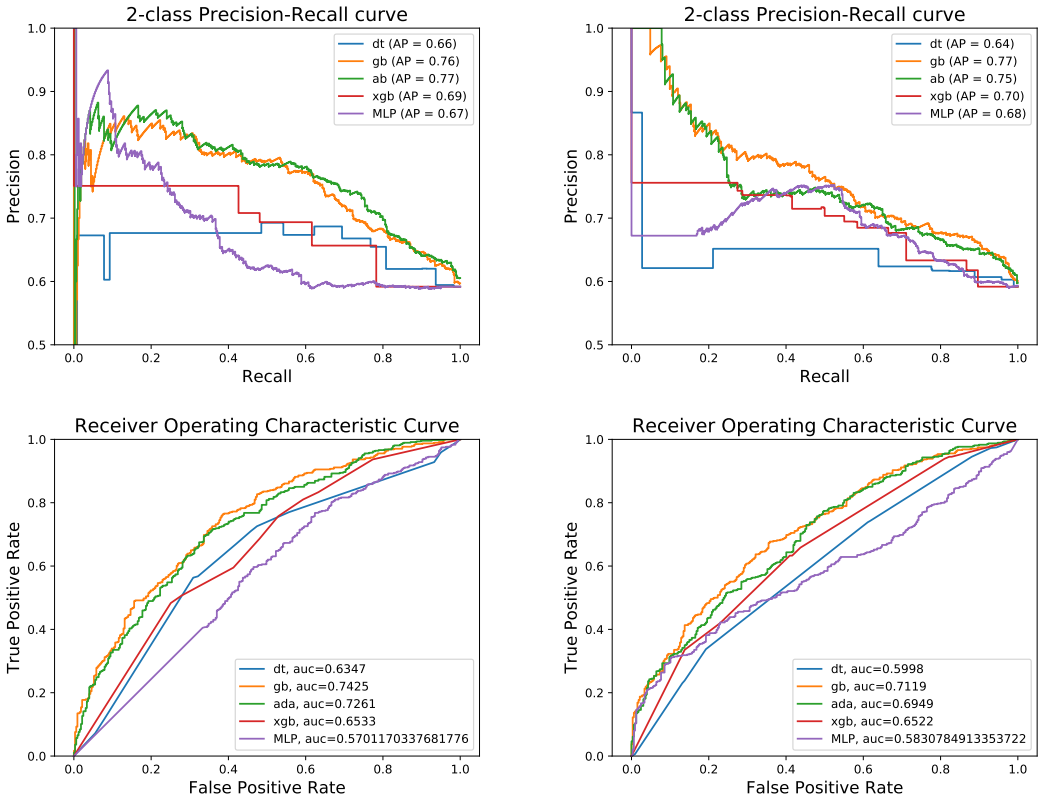

Fig. 9. **Precision-Recall and Receiver Operating Characteristic Curves for the best performing model in accordance to median. We compare the Wiki and APNews Word2Vec embeddings performance with in the Word2Vec, TF-IDF, Rules models here. Precision-Recall Curve for Wiki embeddings (Top Left), Precision-Recall Curve for APNews embeddings (Top Right), Receiver Operating Characteristic Curves for Wiki embeddings (Bottom Left), and Receiver Operating Characteristic Curves for APNews embeddings (Bottom Right).**

amount of information being gathered by the CDC in order to understand side effects and the misinterpretation of side effects could result in more serious issues. Addressing this issue would require a more precise labeling of the data and working with a larger amount of data.

The relevant features for vaccine severity within tweets could be improved. There has been recent work that shows the importance of a very careful feature engineer for a specific domain [16]. Some features could be investigating URLs, emoticons, sentiment of images, engagement scores, investigating profile biography information and others.

XI. Conclusion

By developing ML models with natural language features at the forefront, we developed methods to help government organizations, like the CDC, determine the vaccine side effects that people are experiencing with high accuracy. We developed and benchmarked the first set of ML methods to predict vaccine severity with only information that is available through twitter. This work demonstrates our ability to predict vaccine severity with high accuracy when focusing on natural language features and features showing users influence, i.e. favourites count.

We expect that these methods will be a valuable resource for the CDC by allowing them to understand the sideeffects that patients are receiving from the COVID-19 vaccine. For example, instead of collecting the twitter data and having them labeled they will be able to use our methods for an automatic vaccine severity prediction. Through these methods all parties involved will be able to make a more informed decision about potential side effects that someone could experience. All pre-processing, feature extracting, training, and benchmarking code is publicly available at https://github.com/DevinJMcc/Social_media_mining.

Acknowledgment

We are thankful to Twitter for having open and easy access to the developer platform. Swapna Gokhale for labeling this data and mentoring us throughout the project.

References

[1] Brian P. Dunleavy. Most u.s. teens, young adults want to get covid-19 vaccine, survey finds. *Health News*.

[2] Firoj Alam, Fahim Dalvi, Shaden Shaar, Nadir Durrani, Hamdy Mubarak, Alex Nikolov, Giovanni Da San Martino, Ahmed Abdelali, Hassan Sajjad, Kareem Darwish, et al. Fighting the covid-19 infodemic in social media: a holistic perspective and a call to arms. *arXiv preprint arXiv:2007.07996*, 2020.

[3] Emily Chen, Kristina Lerman, Emilio Ferrara, et al. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273, 2020.

[4] Gautam Kishore Shahi and Durgesh Nandini. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*, 2020.

[5] Umair Qazi, Muhammad Imran, and Ferda Ofli. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15, 2020.

[6] Xinyi Zhou, Apurva Mulay, Emilio Ferrara, and Reza Zafarani. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3205–3212, 2020.

[7] İrfan Aygün, Buket Kaya, and Mehmet Kaya. Aspect based twitter sentiment analysis on vaccination and vaccine types in covid-19 pandemic with deep learning. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2021.

[8] Liviu-Adrian Cotfas, Camelia Delcea, Ioan Roxin, Corina Ioanăş, Dana Simona Gherai, and Federico Tajariol. The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEE Access*, 9:33203–33223, 2021.

[9] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. *Scientific Reports*, 10(1):1–10, 2020.

[10] Cristina M Pulido, Beatriz Villarejo-Carballido, Gisela Redondo-Sama, and Aitor Gómez. Covid-19 infodemic: More retweets for science-based information on coronavirus than for false information. *International Sociology*, 35(4):377–392, 2020.

[11] Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. Detecting east asian prejudice on social media. *arXiv preprint arXiv:2005.03909*, 2020.

[12] Nijhum Paul and Swapna S. Gokhale. Analysis and classification of vaccine dialogue in the coronavirus era. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3220–3227, 2020.

[13] Richard F. Sear, Nicolás Velásquez, Rhys Leahy, Nicholas Johnson Restrepo, Sara El Oud, Nicholas Gabriel, Yonatan Lupu, and Neil F. Johnson. Quantifying covid-19 content in the online health opinion war using machine learning. *IEEE Access*, 8:91886–91893, 2020.

[14] Devin J McConnell, James Zhu, Sachin Pandya, and Derek Aguiar. Case-level prediction of motion outcomes in civil litigation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 99–108, 2021.

[15] Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499, 2015.

[16] Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. Simple open stance classification for rumour analysis. *arXiv preprint arXiv:1708.05286*, 2017.

[17] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] Joseph Lilleberg, Yun Zhu, and Yanqing Zhang. Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140. IEEE, 2015.

[20] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):1–21, 2007.

[21] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[22] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[23] Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.

[24] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[27] Thomas Hancock, Tao Jiang, Ming Li, and John Tromp. Lower bounds on learning decision lists and trees. *Information and Computation*, 126(2):114–122, 1996.

[28] Hyafil Laurent and Ronald L Rivest. Constructing optimal binary decision trees is np-complete. *Information processing letters*, 5(1):15–17, 1976.

[29] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

[30] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[31] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3):1937–1967, 2021.

[32] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated machine learning*, pages 3–33. Springer, Cham, 2019.

[33] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[34] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

[35] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Feature	Description	SF
V/L	Severe or Mild side effects	
Angry	The percent anger seen in the tweets	
Fear	The percent fear seen in the tweets	
Happy	The percent happy seen in the tweets	
Sad	The percent sad seen in the tweets	
Positive	Tweet considered overall positive	
Negative	Tweet considered overall negative	
Neither	Tweet considered overall neutral	
Favourites Count	Number of tweets a user favorites	✓
Friends Count	The number of people a user is following	✓
Followers Count	The number of users following a single user	✓
Statuses Count	Number of tweets	✓
Listed Count	The number of public lists they are apart of	✓
Favorite Count	Number of favorites the tweet has	✓
Retweet Count	Total number of retweets	✓
Location	The geo-cordinates of tweets if active	✓
Verified	If a verified user	✓
is Retweet	If the tweet is a retweet	✓

TABLE IV

DESCRIPTION OF ALL FEATURES WITHIN THE RAW DATA. HERE FEATURES ARE ALL THE FEATURES FROM WITHIN THE RAW DATASET, DESCRIPTIONS ARE FROM OUR ANALYSIS DURING THE LABELING OR THE TWITTER API, AND SF IS FOR WHETHER OR NOT IT IS A SOCIAL FEATURE OR NOT.

	Models				
Data	**dt**	**ada**	**xgb**	**gb**	**MLP**
Database	.58	.59	.553	**.5925**	.58
APNews	.632	.6487	**.657**	.6528	.5863
APNews TF-IDF	.6153	.6133	.6507	**.6632**	.4782
APNews TF-IDF Rules	**.6092**	.5967	.578	.6091	.5925
Wiki	.6341	.6175	.6507	**.659**	.5904
Wiki TF-IDF	.6528	.6237	.6549	**.6819**	.6154
Wiki TF-IDF Rules	**.6487**	.5925	.5842	.6154	.6112

TABLE V

PREDICTING VACCINE SEVERITY FROM TWITTER SOCIAL FEATURES AND TEXT FEATURES. HERE WE HIGHLIGHT THE BEST PERFORMING MODEL FOR EACH OF THE DATASETS CREATED. WE SEE WIKI WORD2VEC EMBEDDINGS WEIGHTED BY TF-IDF HAD THE LARGEST CLASSIFICATION ACCURACY AFTER HYPERPARAMETER OPTIMIZATION. WE ALSO SEE APNEWS WEIGHTED BY TF-IDF WAS THE NEXT HIGHEST ACCURATE MODEL AFTER HYPER PARAMETER OPTIMIZATION. THE CLASSIFICATION MODELS ARE: ADABOOST (ADA), DECISION TREE (DT), XGBOOST(XGB), GRADIENT BOOSTING(GB), MULTILAYER PERCEPTRON (MLP).

	Models				
Data	model	**Accuracy**	**Precision**	**Recall**	**F1**
Database	dt	0.5886	0.5921	**0.9684**	0.7337
	gb	0.5918	0.5918	**1.0**	0.7435
	xgb	0.568	0.5918	**0.8618**	0.7037
	ada	0.5918	0.5918	**0.9979**	0.7426
	MLP	0.5886	0.5918	**0.9852**	0.7377
W2V	dt	0.6161	0.6425	**0.7911**	0.7105
	gb	0.6392	0.6589	**0.8143**	0.7279
	xgb	0.6261	0.6781	**0.6994**	0.6896
	ada	0.6267	0.6220	**0.9335**	0.7481
	MLP	0.5793	0.5956	**0.9325**	0.7241
W2V TF-IDF	dt	0.608	0.6485	**0.7342**	0.6925
	gb	0.6667	0.6717	**0.8513**	0.7511
	xgb	0.6161	**0.6803**	0.6656	0.6724
	ada	0.623	0.6136	**0.9831**	0.755
	MLP	0.60	0.6416	**0.7838**	0.71
W2V TF-IDF Rules	dt	0.6017	0.6462	**0.7479**	0.6861
	gb	0.6067	0.6119	**0.9367**	0.7384
	xgb	0.5780	**0.6792**	0.5401	0.6042
	ada	0.5918	0.5918	**1.0**	0.7435
	MLP	0.5724	0.5948	**.8734**	0.7082

TABLE VI

PREDICTING VACCINE SEVERITY FROM TWITTER SOCIAL FEATURES AND TEXT FEATURES. THIS TABLE IS REPORTING THE MEDIAN ACCURACY FROM THE 100 BOOTSTRAPPED RUNS FROM EACH MODEL. WE SEE THAT FOR ALL ML MODELS THE COMBINED AVERAGE MEDIAN ACCURACY FOR BOTH WORD2VEC,TF-IDF IS HIGHER THAN THE COMBINED AVERAGE MEDIAN FOR OTHER MODELS. WE HIGHLIGHT THE MODEL THAT PERFORMS BEST IN REGARDS TO ALL METRICS AND NOTE THAT OUR BEST PERFORMING MODEL IS W2V, TF-IDF WITH AN GRADIENT BOOSTING CLASSIFIER. THE CLASSIFICATION MODELS ARE: ADABOOST (ADA), DECISION TREE(DT), XGBOOST (XGB), GRADIENT BOOSTING (GB), AND MULTILAYER PERCEPTRON (MLP).

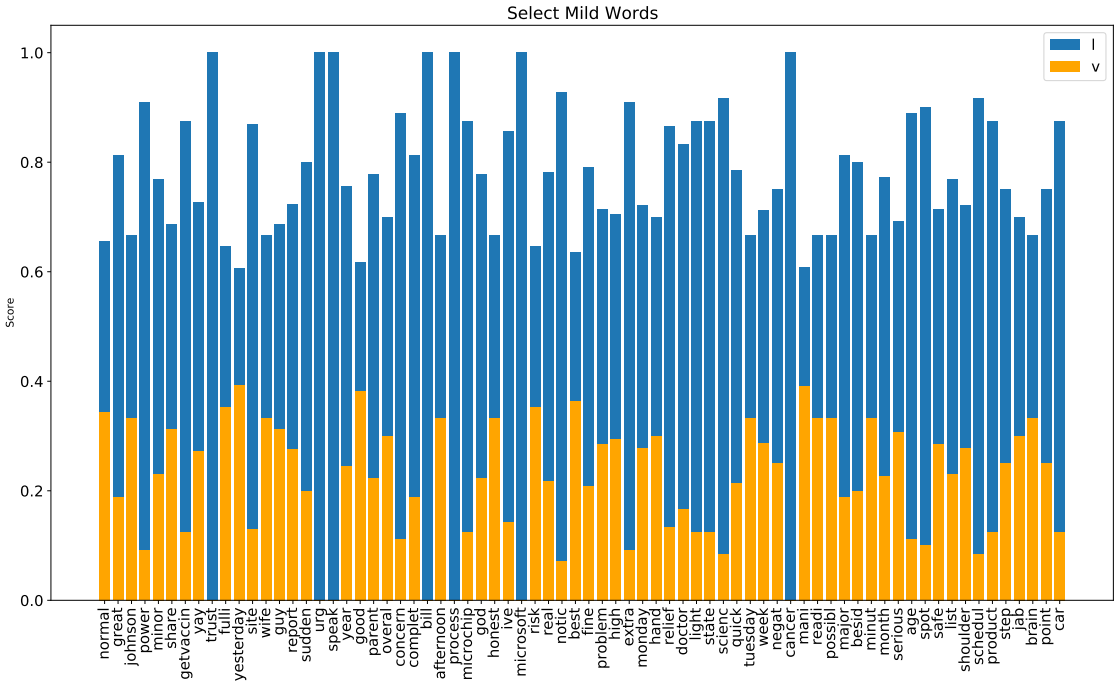

Fig. 10. **Scores For Select Keywords From Mild Tweets. Select keywords from the mild side effects tweets are presented here and scored for there representation in both classes. The way we scored each word is defined under Sec. VI-A2. The differences between scores show us how diverse the opinions are between the two classes. For example, the severe keywords tend to represent pain, discomfort, and medical terms as shown by words like microchip, microsoft, normal, great, etc, which have a higher prevalence in mild tweets.**

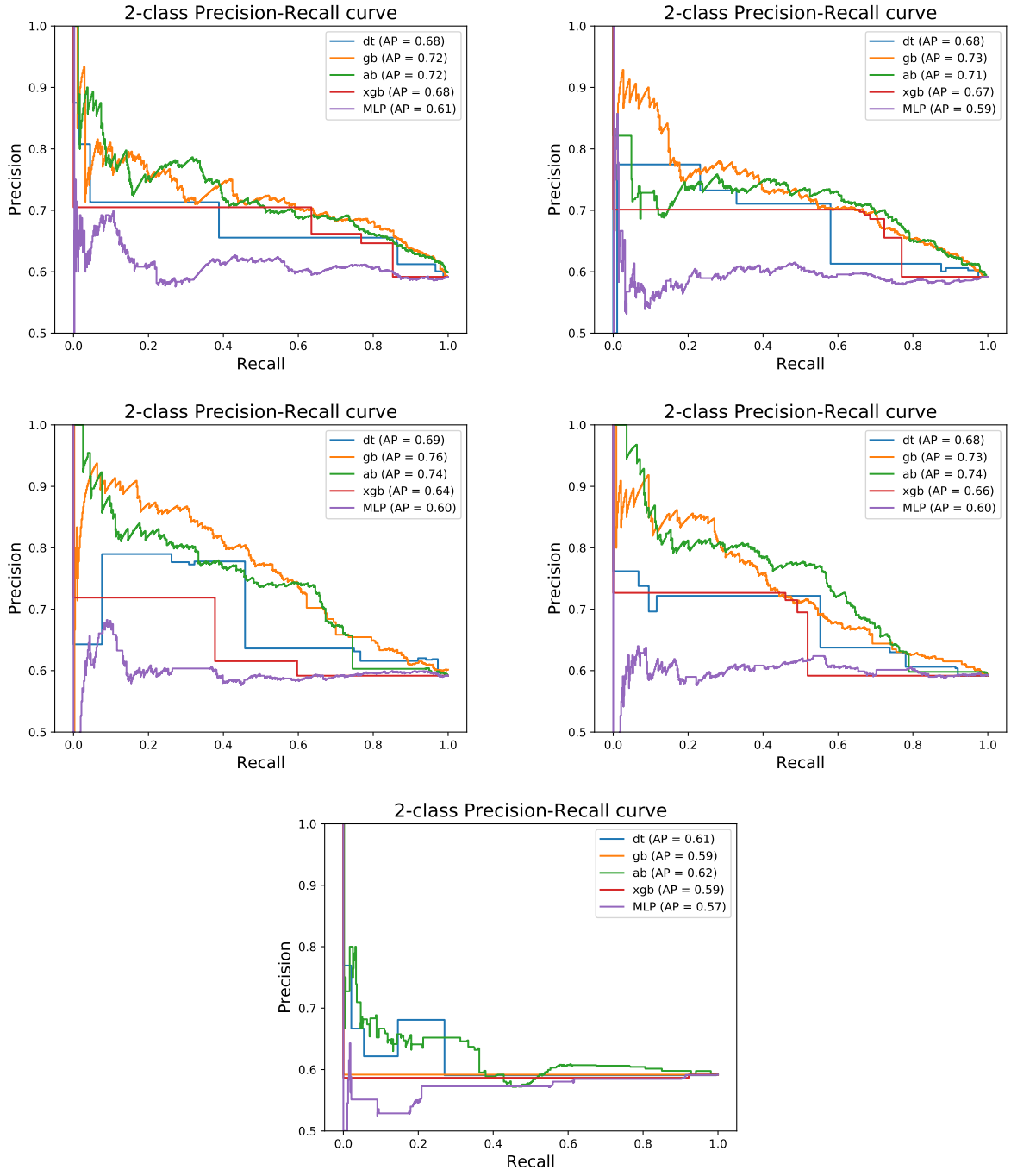

Fig. 11. **Precision-Recall Curves for the remaining models. Precision-Recall Curve for Wiki embeddings (Top Left), Precision-Recall Curve for APNews embeddings (Top Right), Precision-Recall Curve for Wiki embeddings weighted by TF-IDF (Bottom Left), Precision-Recall Curve for APNews embeddings weighted by TF-IDF (Bottom Right), Precision-Recall Curve for database only features (Bottom Center).**

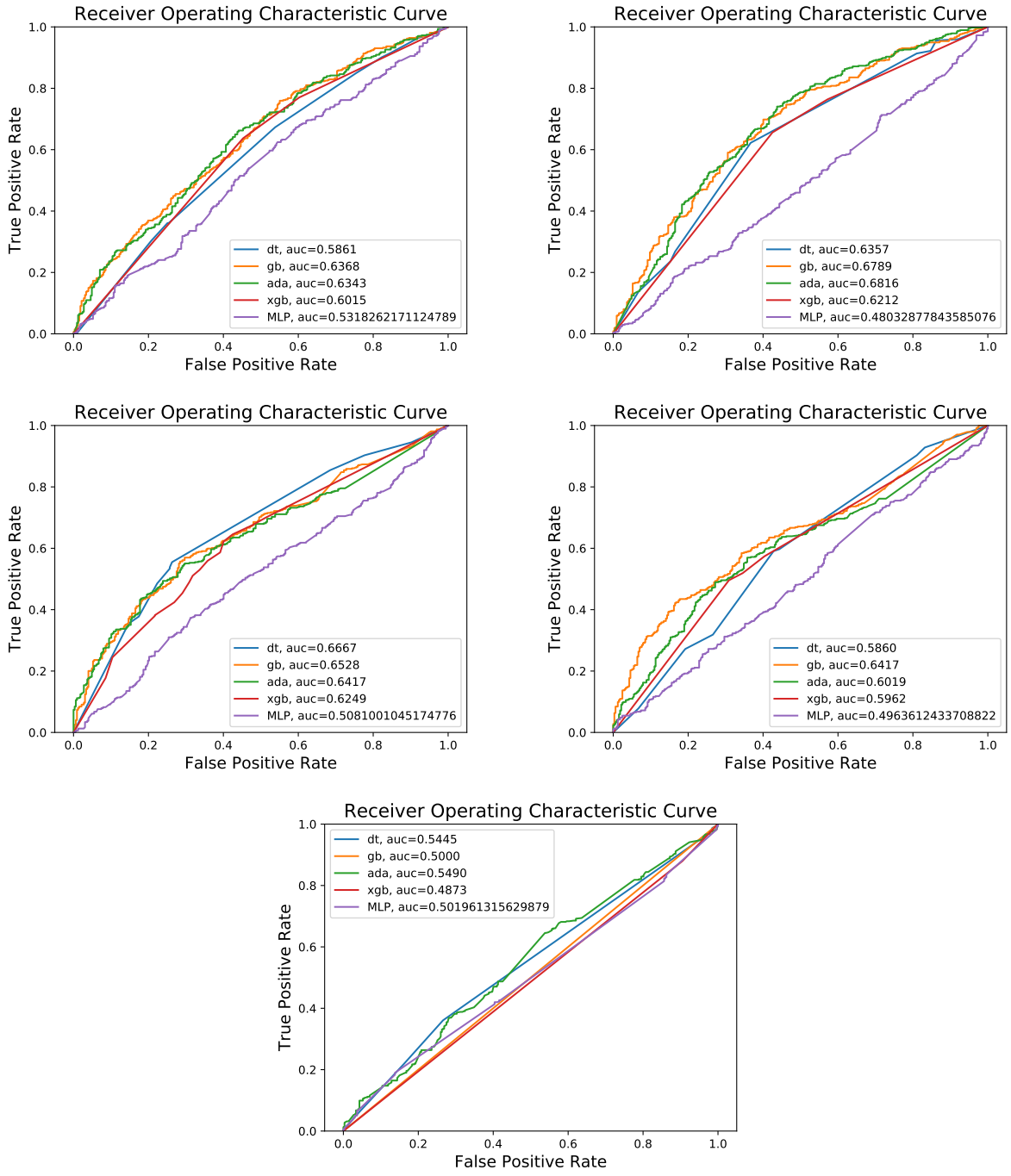

Fig. 12. **Receiver Operating Characteristic Curves for the remaining models. Receiver Operating Characteristic Curve for Wiki embeddings (Top Left), Receiver Operating Characteristic Curve for APNews embeddings (Top Right), Receiver Operating Characteristic Curve for Wiki embeddings weighted by TF-IDF (Bottom Left), Receiver Operating Characteristic Curve for APNews embeddings weighted by TF-IDF (Bottom Right), and Receiver Operating Characteristic Curve for Database only features (Bottom Center).**