# Memorandum

**Date:** March 31, 2019

**To:** Food and Drug Administration (FDA), Adverse Drug Event Monitoring Group

**From:** KS Consulting (Kristi Dunks and Serena Patel)

**Subject:** Detecting High Severity Drug Experiences of Consumers

## 1. Executive Summary

Our analytics team at KS Consulting developed a machine learning text classifier to accurately detect whether a given consumer's drug experience is high severity using key indicators such as "Duration of Days" and text related to high severity identified from sentiment analysis. Our evaluation consisted of ten models that were assessed throughout each phase in the analysis to determine the best model to optimize the FDA's detection of these events. After reviewing our evaluation metrics consisting of accuracy, area under the curve (AUC), recall, precision, and ROC curves, we recommend that the FDA use the Combined CNN and CharCNN model to detect high severity events and use feature relation network (FRN) using support vector machine (SVM) for those cases when mobile devices and the cloud are used for the data. The Combined CNN and CharCNN model had the best performance and effectiveness and therefore, exhibited the most stability and least risk per the results. Use of these models will ensure that the FDA is able to properly classify high and low severity cases at a better rate.

## 2. Background and Description of Analysis

To begin our analysis, we first explored and wrangled structured and unstructured data to construct new features for our models. In doing so, we were able to train and evaluate the model to identify insightful takeaways, such as predictors of high severity, that are described in our analysis below. To understand high severity adverse drug reactions, we used self-reported event texts from public health reporting databases that were categorized as high or low severity. We then developed a machine learning text classifier to understand these texts and predict for high severity events. The FDA provided us with three data files that included textual data which showed the written comments of the adverse events and whether they were rated as high or low severity events (see figure 1), structured data that included the age, gender, average day of stay, and medicine of each reporter, and processed data for FRN that uses various linguistic resources for semantic and syntactic analysis. Each category of data was provided in training, testing, and validation datasets.

**Figure 1. High and low severity Word Clouds**

From the initial textual analysis using Bag of Words, we identified key textual indicators that we then input as dummy variables. These included day, take, doctor, pain, feel, side, anxiety. To provide sentiment to the text, we used the Harvard General Inquirer Dictionary to categorize the words. The dictionary is divided into numerous topic areas and we focused on those related to health, feelings, and positive/negative categories. We also used a happiness dictionary developed by Hedonometer that creates happiness scores for text to further understand the meaning within each report. Within our FRN model using SVM, numerous dictionaries and lexicons had been embedded within the n-gram representations for the text classification tasks. We also used the latent dirichlet allocation (LDA) model to understand the topic distribution within the text and identified the general top topics. The top three topics included (1) hormones, (2) negative side effects/other diseases, and (3) alcoholism/addiction. After reviewing the structured data, the decision tree output highlighted that the 'Duration of Days' variable was the greatest driver of high severity. Therefore, we further explored and enhanced our model by incorporating a dummy variable that encompassed all of the 'Days' with a frequency of 200 or more.

We performed three different analyses, consisting of ten models total, that included training and tuning the models against training data sets and then evaluating the models against testing sets. Specifically, we evaluated each model using 4 metrics: accuracy - percentage of correctly predicted classes out of total predictions, AUC - gauges how accurate the prediction is, recall - ratio of correctly predicted positive classes to all actual positive classes, and precision - ratio of correctly predicted positive classes to all predicted positive classes. First, we tested three models that included SVM, neural networks, and random forest, using textual and structured data separately, and then together. The best results were obtained by combining everything within the models and the accuracy, recall, precision, and AUC are shown in table 1. SVM and neural networks were the top performers, followed by random forest. The ROC curve showed that text or text and structured data perform much better than structured data on its own. This highlights the importance of combining text with structured data in classifying high severity reports. We then analyzed the data using textual, structured, and processed data using FRN, convolutional neural network (CNN), character convolutional network (CharCNN), and Combined CNN and CharCNN. The best results were obtained by combining everything within the models and the associated accuracy and scores are shown in table 1. When combining CNN and CharCNN we obtained the best fit. Therefore, the combined model has lower frequency in alarm failures, where reports slip through the cracks of high severity classification, and false alarms, where reports are incorrectly classified as high severity.

**Table 1. Model performance evaluation metrics (%)**

| | 1. Text, Demographic Models | | | 2. Text Analysis - Deep Learning Models | | | | 3. Supervised Learning Models | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SVM | Neural Networks | Random Forest | FRN (SVM) | CNN | CharCNN | CNN, CharCNN Combined | Decision Tree | Random Forest | XGBoost |
| **Accuracy** | 82.4 | 81.8 | 72.5 | 80.2 | 83.5 | 80.6 | 83.5 | 65.5 | 69.5 | 71.0 |
| **AUC** | 89.8 | 90.0 | 86.2 | 88.0 | 89.5 | 80.8 | 90.5 | 69.8 | 78.3 | 78.2 |
| **Precision Low** | 80.9 | 81.4 | 86.8 | 81.9 | 83.3 | 81.6 | 83.4 | 71.3 | 70.4 | 70.5 |
| **Precision High** | 83.8 | 82.1 | 67.3 | 78.4 | 81.0 | 79.9 | 83.6 | 61.6 | 68.3 | 67.9 |
| **Recall Low** | 82.1 | 79.5 | 49.1 | 80.4 | 83.1 | 82.3 | 85.9 | 57.5 | 72.4 | 71.8 |
| **Recall High** | 82.8 | 83.8 | 93.3 | 80.0 | 81.3 | 79.1 | 80.8 | 74.3 | 66.2 | 66.5 |
| **Skill Score** | N/A | N/A | N/A | N/A | N/A | N/A | N/A | 11.8 | 22.8 | 22.8 |

For our third analysis, we used decision tree, random forest, and XGBoost (see table 1) models. Of these models, random forest performed the best on AUC, followed by XGBoost. For accuracy, XGBoost had the highest score, followed closely by random forest. When assessing these three models and their respective recall and precision scores, the random forest model exhibited the most stability in terms of correctly identifying low and high severity cases.

Moving forward, recurrent assessment of the any changes to the FDA's adverse event report data and new industry advancements to incorporate into the model will be beneficial in continuing to refine the FDA's ability to accurately detect high severity events.

## 3. Key Factors for High Severity

Our analysis highlights that there are several key indicators of high severity events in consumer adverse event reports. The decision tree from our third analysis highlights that duration of days was a primary feature of importance. Other drivers of high severity from our list of feature importance include: happiness score (17%), age (8%), positive and negative (4%) scores, and well-being score (4%). In identifying these key factors for high severity, FDA will have ability to accurately categorize high severity adverse drug reaction reports.

## 4. Recommendation / Summary

The Combined CNN and CharCNN as well as the FRN using SVM models yield the strongest predictive power per the evaluation metrics. Namely, the AUC score for the combined model is 90.5, indicating a higher accuracy in classifying high severity. Both models work well for text data from reports, but have trade-offs. Deployment of the combined model would require model management discussions surrounding computational speed, as this model is scalable to large data sets and can represent complex relations. However, while the combined model performs best in classifying high severity reports, the algorithm requires a full data set, which may lead to potential privacy issues and why we offer the FRN using SVM model as our second option. While the evaluation metrics for the FRN model are not as high as the combined model metrics, the FRN model not only can be run on the cloud but also does not need a full data set, which mitigates potential privacy concerns and reduces the need for greater computational power for processing data. Additionally, this model has high stability and low risk in correctly classifying low and high severity similar to the other deep learning models (see figure 2). The results show minimal variance among those ROC curves based on the different tuning parameters in setting A, C, F curves. Per this recommendation, the FDA can utilize the key drivers of high severity and the combined and FRN models to identify high severity reports precisely with limited risk of misclassification.

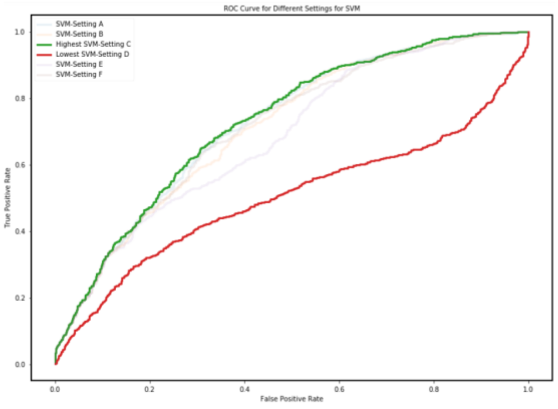

**Figure 2. ROC variance model for FRN SVM model**

Reference: Domino Workbook Link