Catalina Perez,
Isaiah Herard,
Nash Morrison

Project 1: Corona & Triage Dataset Evaluations

Google Colab Notebook

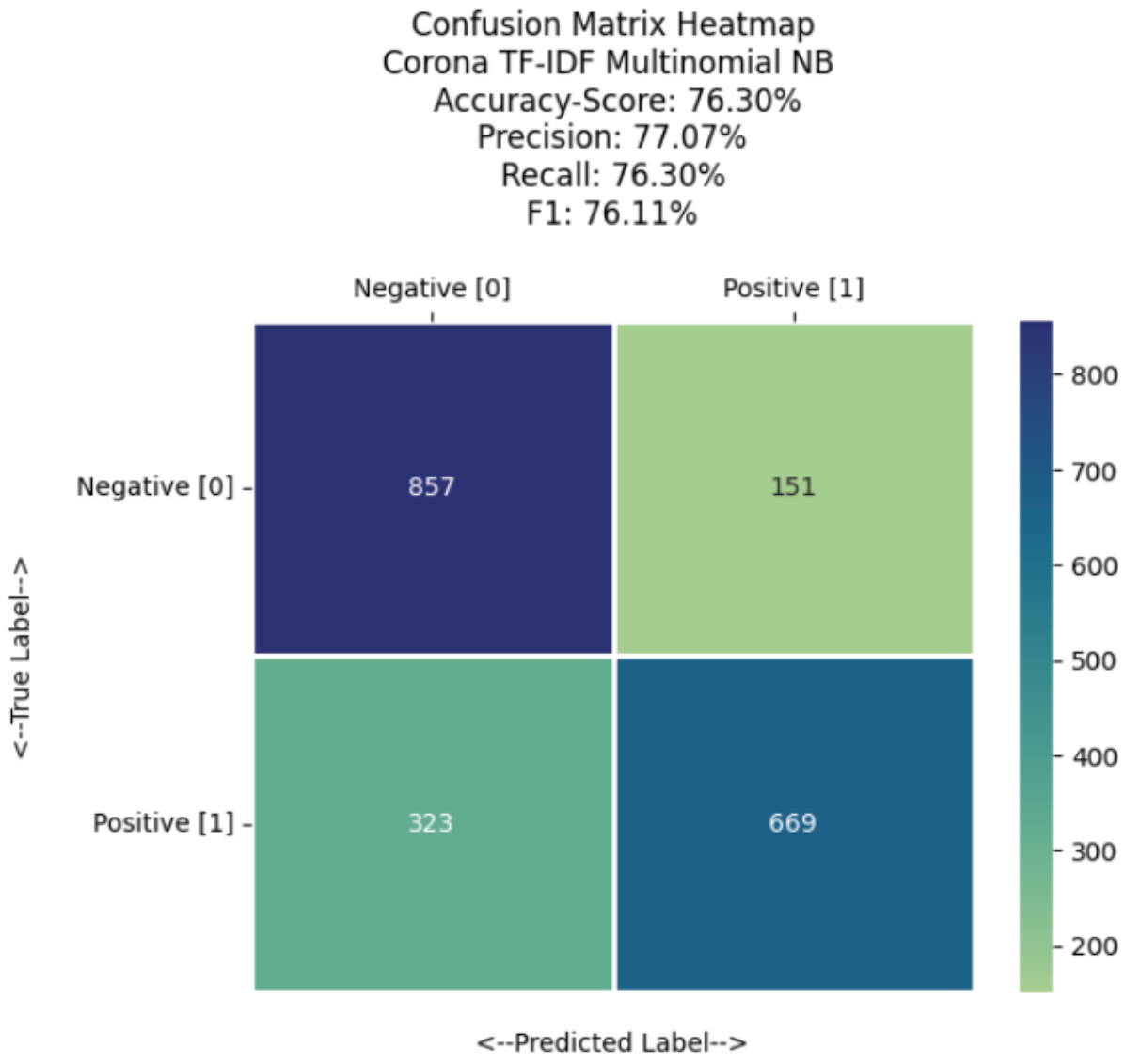
Our Corona and Triage Dataset Evaluation focuses on classifying text data using two different models for machine learning: Multinomial Naïve Bayes, based of Baye's Theorem and Decision Tree Classifiers, tree-like structures that classify data. Both datasets consisted of a collection of text samples, labeled into categories by separation of an | to indicate negative/no aid [0] or positive/aid [1] responses. For transforming our text into numerical expressions, our group went ahead and applied two different vector techniques from our sklearn libraries and used Pandas and Numpy to create our data frames and arrays. The *Count Vectorizer* matrix values show how often each word appears in our document, whereas the *TF-IDF adjusts for word importance* which considers how frequently a word appears within the corpus.

Corona	*Count Vectorizer*			*TF-IDF Vectorizer*		
	Accuracy	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*
Multinomial NB	.7455	.7486	.7455	.7630	.7707	.7630
Decision Tree	.5875	.6610	.5875	.5720	.6428	.5720

Triage	*Count Vectorizer*			*TF-IDF Vectorizer*		
	Accuracy	*Precision*	*Recall*	*Accuracy*	*Precision*	*Recall*
Multinomial NB	.6874	.7065	.6874	.6854	.6927	.6854
Decision Tree	.6738	.7114	.6738	.6718	.7095	.6718

Finalized results show that our best Vectorizer (TF-IDF) pulled a 76% for the Corona data, while the Count vectorizer gave results around 68%. However, TF-IDF provided better overall results for our experimentation because of the adjustments the algorithm makes when training the data.

Catalina Perez,
Isaiah Herard,
Nash Morrison



Catalina Perez,
Isaiah Herard,
Nash Morrison