COVID-19 News Analyzer

Team Name: CRUW

Carl Barbee
crb616

Rahul Vadaga
srv304
Responsible for submission

Udit Arora
ua388

Qingfu Wan
qw894

***Abstract*—We have created a tool, utilizing supervised learning techniques, that lets users analyze news articles related to COVID-19 on a variety of metrics—fakeness, sentiment, emotion, and category. This helps in getting a sense of how different news outlets covered the crisis, affecting the economy as well as the health and well-being of people living through the crisis. We collected and annotated a group of 24 COVID-19 articles on each of these metrics, and report the accuracy of our models on these articles as well as the metric-specific datasets. We investigated whether there has been a rise in negative sentiment such as fear, anger, or disgust, in addition to, the effect of fake news.**

I. Introduction

A. Problem Description

The current pandemic has devastated communities' health and economy at a time where news coverage has been inconsistent depending on the source. Moreover, the deluge of information has made it difficult for any individual to analyze the news coverage. Machine learning combined with natural language processing can provide useful tools for parsing and analyzing this information. Hence it is incumbent upon us to measure how serious and effective our response has been in the context of the havoc—societal, economic, mortality—the virus has caused. Such important aspects can be studied by scrutinizing news media considering they provide essential information to their readership. However, within the past few years, we have also seen a rise in misinformation across the world [1]. Therefore, it would be a good exercise to understand whether media outlets have shifted the narrative away from reality or downplayed the impact.

B. Approach Used

We were mainly interested in analyzing how different outlets covered the pandemic to understand who took things seriously and when. Towards that goal, we utilized a set of classifiers to analyze news articles scraped from the web to determine the emotion, sentiment, fakeness, and category. These classifiers were trained on standard datasets available for each task [2]–[5], since there are no datasets available that are specific to COVID-19 news articles. We collected content from a diverse set of news outlets: Breitbart, CNN, LA Times, The New Yorker, NPR, NY Times, The Guardian, Vox, South China Morning Post, and Wired, to provide our models with a representative dataset. Each metric required a different set of models to yield a good outcome.

C. Summary of Results and Contribution

Our primary contribution is deploying machine learning techniques to analyze trends along different metrics about the novel coronavirus disease (COVID-19). We trained multiple classifiers on standard datasets and applied them to a test set of COVID-19 articles. We found certain metrics yielded better results than others. Namely, accuracy on Fakeness and News Categorization was quite good, but Emotion and Sentiment gave mediocre performance. This is not too revealing given the change in domain and granularity between test articles and the training data. Surprisingly, some simple models such as Logistic Regression performed better than more complex models, e.g. AdaBoost, in certain domains. The results on the test set of COVID-19 articles are shown in Table I. Code is available at the following link.

II. Related Work

The models and techniques employed are quite common in NLP literature [6]–[11]. There has been active engagement within the machine learning community in tackling open research [12], social media [13], transmission rates [14], and emotions [15], among others, to find potential ways to respond to this crisis. Each of these ways utilize a variety of NLP techniques to sift through vast amounts of data to extract and summarize information for medical researchers, policy makers, and everyday citizens. We were particularly interested in Chen's project from NYU [16]. Chen et al. have created an automated data collection and analysis pipeline to assist the public in understanding how misinformation spreads on social media. However, there has been little work done in the domain of news articles related to COVID-19. We found Chen's work fascinating and decided to extend the idea to news outlets to determine not only misinformation, but to also do a comprehensive analysis of the coverage.

III. Problem Definition and Algorithm

A. Task

At a high level, we have 4 tasks—sentiment analysis, fake news detection, emotion detection and news article categorization—which go through a training phase (Fig. 1) and a prediction phase (Fig. 2). In the training phase, for each task (say, Fakeness), we separately train a series of models with a set of feature transformations, using a standard dataset. We then evaluate all the trained models based on their macro

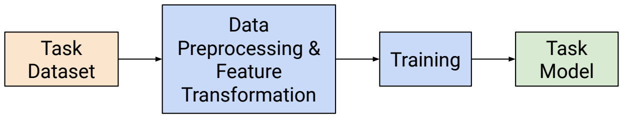


Fig. 1. Training Phase

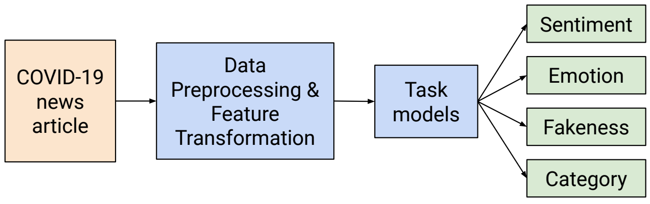


Fig. 2. Prediction Phase

precision score on the corresponding test dataset and choose the best-performing model. Finally, in the prediction phase, the best models take the COVID-19 articles as their inputs and outputs labels for the 4 tasks.

B. Algorithm

We treat each metric as a separate machine learning task with a separate dataset and prediction model. The core components of the pipeline remain the same for each task—feature engineering and model building. We experiment with different approaches for each task and pick the combination that gave the best score in terms of macro precision. This makes the approach modular by design and allows us to add more metrics in a plug-and-play manner.

A few models we trained are: Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), ensemble methods like Random Forests, AdaBoost, and XGBoost. Given the number of models and feature combinations for each task, we added tables showing the entire set of results in the Appendix for reference.

We will demonstrate the process through a simple example (see table III-B). Consider the task of news article categorization where we predict the category of an article from its headline.

Column	Value
Headline	The 7 Best Netflix Shows And Movies Debuting May 2018
Category	ENTERTAINMENT

Training Phase:

- *Case Normalization and Non-alphabet Removal:* All uppercase letters are converted to lowercase letters and all non-alphabetic characters (digits, punctuation, etc.) are removed. In the above example, the headline is converted to 'the best netflix shows and movies debuting may '.
- *Lemmatization:* Next, we use the WordNet lemmatizer from the NLTK library to lemmatize the text from the previous step. Lemmatization[1] refers to removing inflectional endings and converting words to their base/dictionary form. As an example, 'saw' would get converted to 'see' or 'saw' depending on whether 'saw' was used as a verb or a noun. The lemmatized text in our example is 'the best netflix show and movie debuting may'.
- *Feature Transformation:* Now, we need to convert text to features before providing it to our model. For example, CountVectorizer in scikit-learn with n-grams of sizes upto 3. Additionally, we use stop-word removal and limit number of features to 5000, which helps filter out less frequent n-grams and increases training speed. Afterwards, the lemmatized text is converted to 'best netflix movie', which is finally converted to a sparse 5000-dimensional vector with ones for 'best', 'netflix' and 'movie'.
- *Model Training:* Finally, the transformed features are trained on a set of models, e.g. MNB, using cross-validation.

Prediction Phase:

- *Input COVID-19 article:* We input a sample COVID-19 news article.
- *Text Pre-processing and Feature Transformation:* This article would go through the pre-processing steps described in the training phase, which outputs as list of features (either BoW, n-grams, TF-IDF).
- *Prediction:* Finally, we predict the labels for the 4 tasks. For this article, the labels predicted by our models are Fake, 2.11, angry-disgusted and sports.

IV. Experimental Evaluation

A. Data

Dataset Details: Several datasets are available for the four tasks. From these, we selected a popular dataset for each task [2]–[5]. For the sentiment dataset, each phrase is one of 5 classes (0-4), where 0 means the sentiment is negative and 4 means positive. In the fake news dataset, an article can be one of 2 classes—fake or authentic. For emotion detection, the primary emotion in a sentence can be one of 5 classes—angry-disgusted, fearful, happy, sad, or surprised. Finally, for categorization, the label for a news article can be one of these categories—Business, Economy, Crime, Entertainment, Politics, Religion, Science, Sports, Travel or Wellness. For each dataset, we plotted the frequency of each class to determine if there was a class imbalance (see Fig. 3). Sentiment, Categories, and Emotion datasets have an imbalanced class distribution, but Fake News dataset is roughly equally distributed.

Data Pre-processing and Feature Engineering: We applied common NLP pre-processing operations such as tokenizing, stop-word filtering, punctuation removal, and lemmatizing [6]–[11]. We then employed one of the common feature extraction methods: Bag of Words (BoW), BoW with

[1] See this page to learn more about lemmatization.

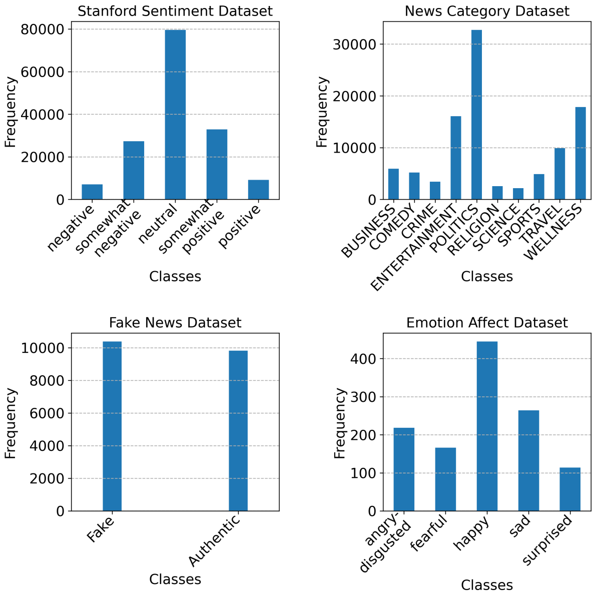


Fig. 3. Training Datasets

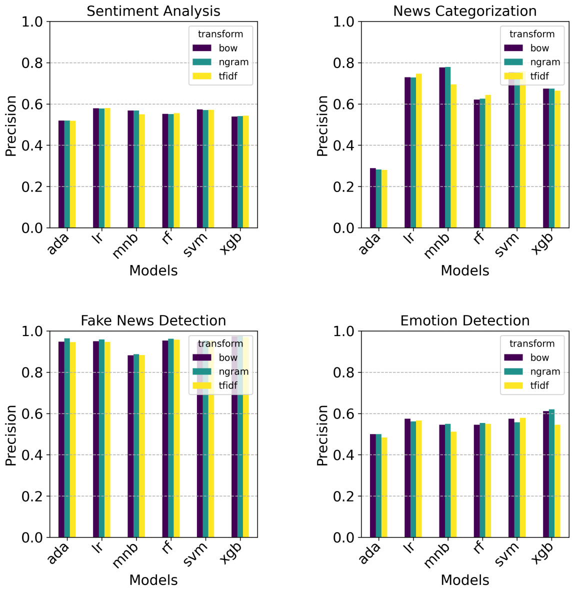


Fig. 4. Experiment Results

Term Frequency (TF-IDF), and BoW with n-grams (up to size 3). We extracted the top 5000 features from each of these transformations and used them to train our models.

B. Methodology

Training Phase: The results of each model and feature transformation are presented both in fig. 4 and the appendix section V-B. We train a series of supervised models and evaluate their validation precision (micro/macro), recall (macro), F1 score (macro), and accuracy. We perform limited hyperparameter tuning for the models using 5-fold cross-validation. We use macro precision to determine the best model using a 80-20 train-test split on the standard datasets, then obtain predictions on the COVID-19 articles. We use macro precision as the comparison metric because we want to maximize the accuracy of our predictions on the COVID-19 articles, irrespective of the class they belong to.

Prediction Phase: While evaluating our models on the COVID-19 articles' test set, we calculate the different metrics for each task. For Fakeness and Categories, we report the accuracy. For Emotion, we analyze the emotions present in different sentences of the articles and report the top-3 classes and calculate the top-3 accuracy. For Sentiment, we calculate the sentiment label for each sentence and take a mean of these values to obtain a final sentiment score. We then calculate the Root Mean Squared Error.

C. Results

Fig. 4 shows the results for each task during the training phase. For Sentiment Analysis, we see similar results for both Logistic Regression and SVM (one-vs-all). In News Categorization, boosting methods show poor performance. This is because our news category dataset has a lot of class imbalance (Fig 3), which means that the samples from some of the minority classes are possibly being treated as outliers, and boosting methods are known to not handle outliers very well. For Fake News Detection, most of the models except for Multinomial Naive Bayes, achieve good results. Finally, for Emotion Detection, we obtained low precision on almost all models because our dataset is relatively small.

TABLE I
Evaluation on COVID-19 News Articles

Best Model / Metric			
Fakeness (Accuracy)	***Sentiment (RMSE)***	***Emotion (Accuracy)***	***Category (Accuracy)***
XGB + ngram	LR + TF	XGB + ngram	NB + ngram
0.708	2.531	0.375	0.583

Table I contains the results from the prediction phase on the 24 COVID-19 news articles that we annotated. We observe that the accuracy on Fakeness and Categorization is satisfactory despite the domain issue, probably on account of the article-level training. The performance on emotion and sentiment dataset shows lower generalization capability compared to Fakeness and Categories. Note, the news category problem is a multi-class classification, unlike its binary counterpart in Fakeness. With that said, an accuracy of 58.3% is fairly high.

D. Discussion

Our main hypothesis was that the standard datasets for each task would translate well to predicting COVID-19 news articles. Considering the Fakeness and News Categorization tasks were trained on datasets containing news content, there was an easier translation to COVID-19 articles. The only concern was that health-related topics would not be covered as frequently in the standard datasets as in the COVID-19 ones. Our models perform reasonably well though.

The paucity of the Emotion dataset (∼1k training samples) leads to moderate training precision as seen in Appendix Table A. The size of the dataset has lead to significant over-fitting in our models. Additionally, the majority of the samples are drawn from books/novels. Novelists are less likely to focus on healthcare or economic data that are quite common in the press. The difference in domain along with translation from sentences to articles helps explain the results on this task partially.

The Stanford sentiment dataset is a large-scale dataset, but our models have mediocre training precision as seen in Appendix A. This suggests the models may be under-fitting. To enlarge the hypothesis function space we also tried to train a deep RNN model, but failed to achieve better results. We speculate this may be caused by applying models trained on phrases to sentences, i.e., phrases and sentences are not aligned. The domain issue is more prevalent for sentiment because the training set is comprised of movie reviews, while we are dealing with news. Further investigation on per-sentence evaluation suggests that the model predicts a mostly neutral sentiment. This is expected as sentiment is a fine-grained task.

V. Conclusion

In this work, we trained a variety of models on standard datasets to perform 4 different tasks relevant to our analysis of COVID-19 news articles. We analyzed the performance of our models on the standard datasets as well as a group of 24 COVID-19 related articles that we annotated. We further developed a web portal for easy access to our models and presented analysis of the trends observed in articles over time.

Some major shortcomings were in training on standard datasets that are not COVID-19 or at least health/news specific. The Emotion and Sentiment datasets contain literary and movie reviews respectively. The dearth of medical topics in these datasets make it difficult to translate to COVID-19 topics—health, economic, politics, etc. Finding a more representative training set for each task would improve our models' results. This is easily seen in Fakeness where it was trained on a Fake News dataset and yielded good results.

Labelling is another area for improvement. We had one team member label the COVID-19 articles for each task. Having multiple members label and finding the common label for each sample, as well as annotating more articles would provide less bias. Exploring a multi-label approach for some tasks might work as well. For example, the News Category samples could have multiple labels such as Politics, Health, and Economic to allow for more nuance.

The simplicity of our models / training process relative to the tasks is another shortcoming in our approach. Utilizing neural networks would potentially yield better results for each task, since the articles are nuanced. We were limited on compute, so we needed to trim down on our cross-validation and hyper parameter tuning which are obvious next steps. Spending more time on model selection and parameter tuning would be helpful.

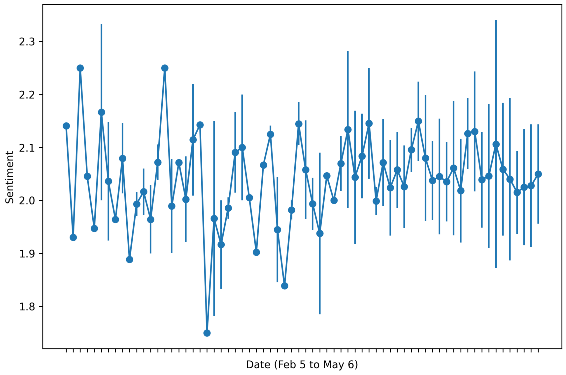
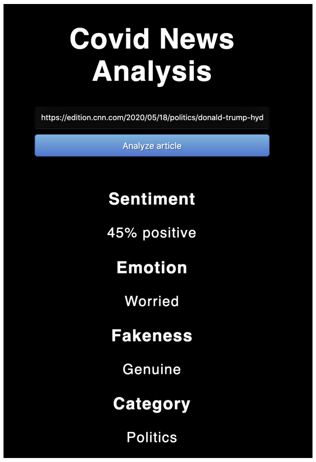


Fig. 5. Sentiment over Time



Covid News Analysis

https://edition.cnn.com/2020/05/18/politics/donald-trump-hyd

Sentiment

45% positive

Emotion

Worried

Fakeness

Genuine

Category

Politics

Fig. 6. Web Portal

A. Historical Analysis

We perform a historical analysis of around 1000 COVID-19 articles on the sentiment analysis task. Figure 5 is a plot of sentiment for published articles from January to May of 2020 where we see the sentiment's variance decreases overtime as expected. In future work, we'll collect more articles and analyze trends for each task.

B. Web Portal

We develop a web portal which lets users analyze any COVID-19 article of their choice by running it against the models trained for each tasks. Figure 6 is a screenshot of the portal in action showing the analysis of an article based on each of the four metrics. If we continue this project, we will collect more COVID-19 articles to improve the accuracy on the four metrics, so that anyone can provide a news article's URL for analysis.

References

[1] D.-M. Ordway, “Fake news and the spread of misinformation: A research roundup,” *Journalist’s Resource*, 2017. [Online]. Available: https://journalistsresource.org/studies/society/internet/fake-news-conspiracy-theories-journalism-research/

[2] E. C. O. Alm, *Affect in text and speech*. Citeseer, 2008.

[3] “Fake news dataset,” 2018. [Online]. Available: https://www.kaggle.com/c/fake-news

[4] R. Misra, “News category dataset,” 06 2018. [Online]. Available: https://www.kaggle.com/rmisra/news-category-dataset

[5] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[6] C. D. Manning, C. D. Manning, and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[7] V. Keselj, “Speech and language processing daniel jurafsky and james h. martin (stanford university and university of colorado at boulder) pearson prentice hall, 2009, xxxi+ 988 pp; hardbound, isbn 978-0-13-187321-6, 115.00,” 2009.

[8] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[9] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[10] H. Daumé III, “A course in machine learning,” *Publisher, ciml. info*, vol. 5, p. 69, 2012.

[11] B. Santorini, “Part-of-speech tagging guidelines for the penn treebank project,” 1990.

[12] E. Zhang, N. Gupta, R. Nogueira, K. Cho, and J. Lin, “Rapidly deploying a neural search engine for the covid-19 open research dataset: Preliminary thoughts and lessons learned,” *arXiv preprint arXiv:2004.05125*, 2020.

[13] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoli, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, “The covid-19 social media infodemic,” *arXiv preprint arXiv:2003.05004*, 2020.

[14] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, “Modelling transmission and control of the covid-19 pandemic in australia,” *arXiv preprint arXiv:2003.10218*, 2020.

[15] B. Kleinberg, I. van der Vegt, and M. Mozes, “Measuring emotions in the covid-19 real world worry dataset,” *arXiv preprint arXiv:2004.04225*, 2020.

[16] Z. Chen, “How does misinformation tracer work?” Mar 2020. [Online]. Available: https://zhouhanc.github.io/cross-platform/

Appendices

Appendix A
Sentiment Analysis Results

	model	transform	precision	precision (macro)	recall (macro)	f1-score (macro)	accuracy
0	mnb	bow	0.568006	0.483012	0.358318	0.382007	0.568006
1	mnb	ngram	0.568455	0.483377	0.358991	0.382963	0.568455
2	mnb	tfidf	0.549598	0.498154	0.284205	0.279198	0.549598
3	svm	bow	0.573073	0.490563	0.343105	0.358390	0.573073
4	svm	ngram	0.570636	0.502908	0.342408	0.358810	0.570636
5	svm	tfidf	0.571502	0.492138	0.352997	0.376456	0.571502
6	lr	bow	0.579359	0.485289	**0.392195**	**0.416603**	**0.579359**
7	lr	ngram	0.578044	0.484621	0.389537	0.414076	0.578044
8	lr	tfidf	**0.580578**	**0.521308**	0.374069	0.402164	0.580578
9	xgb	bow	0.539399	0.478321	0.293735	0.300536	0.539399
10	xgb	ngram	0.541099	0.492681	0.299512	0.308925	0.541099
11	xgb	tfidf	0.542927	0.486080	0.304550	0.316774	0.542927
12	rf	bow	0.552131	0.446020	0.336173	0.354005	0.552131
13	rf	ngram	0.551361	0.451040	0.334670	0.352417	0.551361
14	rf	tfidf	0.555466	0.446703	0.335841	0.351484	0.555466
15	ada	bow	0.519066	0.443485	0.262714	0.251574	0.519066
16	ada	ngram	0.519066	0.443485	0.262714	0.251574	0.519066
17	ada	tfidf	0.518842	0.445653	0.263075	0.252346	0.518842

Appendix B
Emotion Results

	model	transform	precision	precision (macro)	recall (macro)	f1-score (macro)	accuracy
0	mnb	bow	0.545455	0.461286	0.426319	0.431953	0.545455
1	mnb	ngram	0.549587	0.471302	0.428516	0.435708	0.549587
2	mnb	tfidf	0.512397	0.495752	0.348956	0.341614	0.512397
3	svm	bow	0.574380	0.587808	0.456703	0.482489	0.574380
4	svm	ngram	0.557851	0.564264	0.432418	0.452647	0.557851
5	svm	tfidf	0.578512	0.557945	0.450879	0.470985	0.578512
6	lr	bow	0.574380	0.531284	0.481978	0.495037	0.574380
7	lr	ngram	0.561983	0.517858	0.464780	0.476625	0.561983
8	lr	tfidf	0.566116	0.578233	0.429615	0.447860	0.566116
9	xgb	bow	0.611570	0.584485	0.526154	0.535127	0.611570
10	xgb	ngram	**0.619835**	**0.604206**	**0.533297**	**0.543992**	**0.619835**
11	xgb	tfidf	0.545455	0.522488	0.455385	0.472683	0.545455
12	rf	bow	0.545455	0.485145	0.446813	0.454617	0.545455
13	rf	ngram	0.553719	0.514549	0.469670	0.476948	0.553719
14	rf	tfidf	0.549587	0.527377	0.449066	0.465213	0.549587
15	ada	bow	0.500000	0.439407	0.363462	0.369801	0.500000
16	ada	ngram	0.500000	0.439407	0.363462	0.369801	0.500000
17	ada	tfidf	0.483471	0.471146	0.364176	0.370478	0.483471

Appendix C
Fake News Results

	model	transform	precision	precision (macro)	recall (macro)	f1-score (macro)	accuracy
0	mnb	bow	0.881974	0.884401	0.881323	0.881631	0.881974
1	mnb	ngram	0.886933	0.889658	0.886251	0.886580	0.886933
2	mnb	tfidf	0.882718	0.887757	0.881783	0.882125	0.882718
3	svm	bow	0.945450	0.945471	0.945592	0.945447	0.945450
4	svm	ngram	0.954376	0.954339	0.954413	0.954368	0.954376
5	svm	tfidf	0.953880	0.953950	0.953802	0.953861	0.953880
6	lr	bow	0.950161	0.950167	0.950127	0.950145	0.950161
7	lr	ngram	0.958592	0.958686	0.958502	0.958573	0.958592
8	lr	tfidf	0.946442	0.946421	0.946435	0.946428	0.946442
9	xgb	bow	0.974461	0.974496	0.974417	0.974452	0.974461
10	xgb	ngram	**0.977188**	**0.977159**	**0.977215**	**0.977183**	**0.977188**
11	xgb	tfidf	0.969750	0.969720	0.969774	0.969743	0.969750
12	rf	bow	0.953137	0.953710	0.952881	0.953091	0.953137
13	rf	ngram	0.962063	0.962064	0.962040	0.962052	0.962063
14	rf	tfidf	0.958096	0.958440	0.957907	0.958065	0.958096
15	ada	bow	0.947930	0.947939	0.947890	0.947913	0.947930
16	ada	ngram	0.964790	0.964754	0.964855	0.964785	0.964790
17	ada	tfidf	0.945698	0.945850	0.945572	0.945668	0.945698

Appendix D
News Category Results

	model	transform	precision	precision (macro)	recall (macro)	f1-score (macro)	accuracy
0	mnb	bow	0.777088	0.508297	0.573100	**0.523504**	0.777088
1	mnb	ngram	**0.779621**	0.505211	**0.574841**	0.519337	**0.779621**
2	mnb	tfidf	0.694967	**0.603499**	0.423908	0.447800	0.694967
3	svm	bow	0.753838	0.562894	0.486727	0.490846	0.753838
4	svm	ngram	0.754434	0.560219	0.491556	0.492226	0.754434
5	svm	tfidf	0.758706	0.554128	0.506577	0.502710	0.758706
6	lr	bow	0.729296	0.488520	0.519539	0.490947	0.729296
7	lr	ngram	0.728551	0.488457	0.519642	0.490542	0.728551
8	lr	tfidf	0.746684	0.559625	0.508836	0.511782	0.746684
9	xgb	bow	0.674152	0.547751	0.489279	0.490716	0.674152
10	xgb	ngram	0.674052	0.549342	0.484209	0.489602	0.674052
11	xgb	tfidf	0.664613	0.547944	0.465419	0.478187	0.664613
12	rf	bow	0.621044	0.426035	0.424029	0.400052	0.621044
13	rf	ngram	0.625714	0.420159	0.429811	0.400028	0.625714
14	rf	tfidf	0.644443	0.472518	0.407206	0.404894	0.644443
15	ada	bow	0.288688	0.451160	0.196817	0.256991	0.288688
16	ada	ngram	0.282875	0.449648	0.194064	0.254121	0.282875
17	ada	tfidf	0.280640	0.452401	0.190653	0.251818	0.280640