6. Appendix

Appendix 1. Detailed description of classification

We trained classifiers based on three distinct reprocessing steps: 1. Using all words; 2. Removing stopwords 3. Using only the lead text (first 75 words of a news item). No stemming or lemmatization was conducted. We use random sampling to split our dataset into training data (80 percent), on which we trained the three different pre-processing steps, and testing data (20 percent), on which we tested the performance of our final classifiers. We used these steps, as well as Pyhon's scikit-learn machine learning library, to develop a pipeline through which 6 algorithms (Bernoulli Negative Binomial, Multinomial Negative Binomial, Logistic Regression, Stochastic Gradient Descent (SGD), Support Vector Machine Classifier (SVM), Passive Aggressive) were trained on the provided data, constructing three models for each algorithm based on the three text pre-processing options (one using all words, one with stopwords removed, and one using the lead text). This resulted in 18 classifiers.

Additionally, we underwent extensive hyperparameter tuning to ensure best results. For each classifier, we tested different combinations of hyperparameters, including how to use the CountVectorizer function to make the matrix of token counts of text documents in our training corpus; whether or not to use the TfidfTransformer function to convert a count matrix to a normalized tf or tf-idf representation; to establish the number of times the algorithm should iterate over the training texts with optimization iteration; as well as what hinge-loss function to apply. A pipeline was used to establish the best combination of these parameters for each of the 18 classifiers.

Finally, the trained classifiers were tested on the trained data, generating a report for each that showcased each model's performance through its precision, recall, accuracy, and f1 score. With this information, we proceeded to select the best-performing classifier: in this case, SVM with the full text.

Appendix 2. Performance of Supervised Machine Learning Classifiers Tested

Classifier	Preprocessing	Class	Precision	Recall	F1 Score	Support
Passive Aggressive	Full Text	Culture	0.81	0.87	0.83	67.0
Passive Aggressive	Full Text	Disasters and Crimes	0.82	0.79	0.8	67.0
Passive Aggressive	Full Text	Economy	0.88	0.75	0.81	28.0
Passive Aggressive	Full Text	Other	0.88	0.67	0.76	33.0
Passive Aggressive	Full Text	Politics	0.88	0.95	0.91	164.0
Passive Aggressive	Full Text	Sports	0.89	0.83	0.86	41.0
Passive Aggressive	Full Text	Macro Avg	0.86	0.81	0.83	400.0
Passive Aggressive	Full Text	Weighted Avg	0.86	0.86	0.86	400.0
Passive Aggressive	Stopwords Removed	Culture	0.82	0.84	0.83	67.0
Passive Aggressive	Stopwords Removed	Disasters and Crimes	0.76	0.82	0.79	67.0
Passive Aggressive	Stopwords Removed	Economy	0.81	0.79	0.8	28.0
Passive Aggressive	Stopwords Removed	Other	0.91	0.61	0.73	33.0
Passive Aggressive	Stopwords Removed	Politics	0.89	0.94	0.91	164.0
Passive Aggressive	Stopwords Removed	Sports	0.87	0.8	0.84	41.0
Passive Aggressive	Stopwords Removed	Macro Avg	0.84	0.8	0.82	400.0
Passive Aggressive	Stopwords Removed	Weighted Avg	0.85	0.85	0.85	400.0
Passive Aggressive	Only Lead (First 75 words)	Culture	0.74	0.78	0.76	67.0
Passive Aggressive	Only Lead (First 75 words)	Disasters and Crimes	0.8	0.72	0.76	67.0
Passive Aggressive	Only Lead (First 75 words)	Economy	0.83	0.71	0.77	28.0
Passive Aggressive	Only Lead (First 75 words)	Other	0.88	0.64	0.74	33.0
Passive Aggressive	Only Lead (First 75 words)	Politics	0.83	0.92	0.87	164.0
Passive Aggressive	Only Lead (First 75 words)	Sports	0.92	0.88	0.9	41.0
Passive Aggressive	Only Lead (First 75 words)	Macro Avg	0.83	0.77	0.8	400.0
Passive Aggressive	Only Lead (First 75 words)	Weighted Avg	0.82	0.82	0.82	400.0
Bernoulli Negative Binomial	Full Text	Culture	0.81	0.69	0.74	67.0
Bernoulli Negative Binomial	Full Text	Disasters and Crimes	0.81	0.75	0.78	67.0
Bernoulli Negative Binomial	Full Text	Economy	0.9	0.68	0.78	28.0

Appendix 2. Performance of Supervised Machine Learning Classifiers Tested

Classifier	Preprocessing	Class	Precision	Recall	F1 Score	Support
Bernoulli Negative Binomial	Full Text	Other	0.58	0.64	0.61	33.0
Bernoulli Negative Binomial	Full Text	Politics	0.79	0.91	0.85	164.0
Bernoulli Negative Binomial	Full Text	Sports	0.89	0.76	0.82	41.0
Bernoulli Negative Binomial	Full Text	Macro Avg	0.8	0.74	0.76	400.0
Bernoulli Negative Binomial	Full Text	Weighted Avg	0.8	0.79	0.79	400.0
Bernoulli Negative Binomial	Stopwords Removed	Culture	0.82	0.7	0.76	67.0
Bernoulli Negative Binomial	Stopwords Removed	Disasters and Crimes	0.82	0.75	0.78	67.0
Bernoulli Negative Binomial	Stopwords Removed	Economy	0.9	0.68	0.78	28.0
Bernoulli Negative Binomial	Stopwords Removed	Other	0.6	0.64	0.62	33.0
Bernoulli Negative Binomial	Stopwords Removed	Politics	0.79	0.92	0.85	164.0
Bernoulli Negative Binomial	Stopwords Removed	Sports	0.86	0.76	0.81	41.0
Bernoulli Negative Binomial	Stopwords Removed	Macro Avg	0.8	0.74	0.77	400.0
Bernoulli Negative Binomial	Stopwords Removed	Weighted Avg	0.8	0.8	0.8	400.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Culture	0.81	0.81	0.81	67.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Disasters and Crimes	0.8	0.72	0.76	67.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Economy	0.75	0.64	0.69	28.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Other	0.95	0.55	0.69	33.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Politics	0.78	0.93	0.85	164.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Sports	0.92	0.8	0.86	41.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Macro Avg	0.83	0.74	0.78	400.0
Bernoulli Negative Binomial	Only Lead (First 75 words)	Weighted Avg	0.81	0.81	0.8	400.0
Multinomial Negative Binomial	Full Text	Culture	0.82	0.81	0.81	67.0
Multinomial Negative Binomial	Full Text	Disasters and Crimes	0.81	0.69	0.74	67.0
Multinomial Negative Binomial	Full Text	Economy	0.88	0.75	0.81	28.0
Multinomial Negative Binomial	Full Text	Other	0.94	0.48	0.64	33.0
Multinomial Negative Binomial	Full Text	Politics	0.78	0.96	0.86	164.0
Multinomial Negative Binomial	Full Text	Sports	0.91	0.78	0.84	41.0

Appendix 2. Performance of Supervised Machine Learning Classifiers Tested

Classifier	Preprocessing	Class	Precision	Recall	F1 Score	Support
Multinomial Negative Binomial	Full Text	Macro Avg	0.86	0.74	0.78	400.0
Multinomial Negative Binomial	Full Text	Weighted Avg	0.83	0.81	0.81	400.0
Multinomial Negative Binomial	Stopwords Removed	Culture	0.81	0.81	0.81	67.0
Multinomial Negative Binomial	Stopwords Removed	Disasters and Crimes	0.81	0.72	0.76	67.0
Multinomial Negative Binomial	Stopwords Removed	Economy	0.91	0.75	0.82	28.0
Multinomial Negative Binomial	Stopwords Removed	Other	0.95	0.55	0.69	33.0
Multinomial Negative Binomial	Stopwords Removed	Politics	0.79	0.95	0.86	164.0
Multinomial Negative Binomial	Stopwords Removed	Sports	0.91	0.78	0.84	41.0
Multinomial Negative Binomial	Stopwords Removed	Macro Avg	0.86	0.76	0.8	400.0
Multinomial Negative Binomial	Stopwords Removed	Weighted Avg	0.83	0.82	0.82	400.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Culture	0.84	0.85	0.84	67.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Disasters and Crimes	0.81	0.76	0.78	67.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Economy	0.73	0.68	0.7	28.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Other	1.0	0.52	0.68	33.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Politics	0.81	0.93	0.87	164.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Sports	0.92	0.85	0.89	41.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Macro Avg	0.85	0.77	0.79	400.0
Multinomial Negative Binomial	Only Lead (First 75 words)	Weighted Avg	0.84	0.83	0.83	400.0
Logistic Regression	Full Text	Culture	0.74	0.87	0.8	67.0
Logistic Regression	Full Text	Disasters and Crimes	0.81	0.81	0.81	67.0
Logistic Regression	Full Text	Economy	0.7	0.93	0.8	28.0
Logistic Regression	Full Text	Other	0.88	0.7	0.78	33.0
Logistic Regression	Full Text	Politics	0.94	0.9	0.92	164.0
Logistic Regression	Full Text	Sports	0.86	0.76	0.81	41.0
Logistic Regression	Full Text	Macro Avg	0.82	0.82	0.82	400.0
Logistic Regression	Full Text	Weighted Avg	0.86	0.85	0.85	400.0
Logistic Regression	Stopwords Removed	Culture	0.74	0.91	0.82	67.0

Appendix 2. Performance of Supervised Machine Learning Classifiers Tested

Classifier	Preprocessing	Class	Precision	Recall	F1 Score	Support
Logistic Regression	Stopwords Removed	Disasters and Crimes	0.76	0.82	0.79	67.0
Logistic Regression	Stopwords Removed	Economy	0.72	1.0	0.84	28.0
Logistic Regression	Stopwords Removed	Other	0.95	0.61	0.74	33.0
Logistic Regression	Stopwords Removed	Politics	0.96	0.87	0.91	164.0
Logistic Regression	Stopwords Removed	Sports	0.92	0.83	0.87	41.0
Logistic Regression	Stopwords Removed	Macro Avg	0.84	0.84	0.83	400.0
Logistic Regression	Stopwords Removed	Weighted Avg	0.87	0.85	0.85	400.0
Logistic Regression	Only Lead (First 75 words)	Culture	0.69	0.85	0.76	67.0
Logistic Regression	Only Lead (First 75 words)	Disasters and Crimes	0.72	0.78	0.75	67.0
Logistic Regression	Only Lead (First 75 words)	Economy	0.75	0.86	0.8	28.0
Logistic Regression	Only Lead (First 75 words)	Other	1.0	0.58	0.73	33.0
Logistic Regression	Only Lead (First 75 words)	Politics	0.93	0.87	0.9	164.0
Logistic Regression	Only Lead (First 75 words)	Sports	0.93	0.9	0.91	41.0
Logistic Regression	Only Lead (First 75 words)	Macro Avg	0.84	0.81	0.81	400.0
Logistic Regression	Only Lead (First 75 words)	Weighted Avg	0.85	0.83	0.83	400.0
Stochastic Gradient Descent (SGD)	Full Text	Culture	0.77	0.87	0.82	67.0
Stochastic Gradient Descent (SGD)	Full Text	Disasters and Crimes	0.82	0.81	0.81	67.0
Stochastic Gradient Descent (SGD)	Full Text	Economy	0.71	0.96	0.82	28.0
Stochastic Gradient Descent (SGD)	Full Text	Other	0.85	0.67	0.75	33.0
Stochastic Gradient Descent (SGD)	Full Text	Politics	0.95	0.9	0.92	164.0
Stochastic Gradient Descent (SGD)	Full Text	Sports	0.9	0.85	0.88	41.0
Stochastic Gradient Descent (SGD)	Full Text	Macro Avg	0.83	0.84	0.83	400.0
Stochastic Gradient Descent (SGD)	Full Text	Weighted Avg	0.87	0.86	0.86	400.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Culture	0.8	0.84	0.82	67.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Disasters and Crimes	0.77	0.84	0.8	67.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Economy	0.72	1.0	0.84	28.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Other	0.87	0.61	0.71	33.0

Appendix 2. Performance of Supervised Machine Learning Classifiers Tested

Classifier	Preprocessing	Class	Precision	Recall	F1 Score	Support
Stochastic Gradient Descent (SGD)	Stopwords Removed	Politics	0.96	0.9	0.93	164.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Sports	0.88	0.88	0.88	41.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Macro Avg	0.83	0.84	0.83	400.0
Stochastic Gradient Descent (SGD)	Stopwords Removed	Weighted Avg	0.87	0.86	0.86	400.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Culture	0.76	0.79	0.77	67.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Disasters and Crimes	0.71	0.79	0.75	67.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Economy	0.63	0.93	0.75	28.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Other	0.96	0.67	0.79	33.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Politics	0.94	0.86	0.9	164.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Sports	0.93	0.93	0.93	41.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Macro Avg	0.82	0.83	0.81	400.0
Stochastic Gradient Descent (SGD)	Only Lead (First 75 words)	Weighted Avg	0.85	0.83	0.84	400.0
Support Vector Machine (SVM)	Full Text	Culture	0.73	0.88	0.8	67.0
Support Vector Machine (SVM)	Full Text	Disasters and Crimes	0.81	0.84	0.82	67.0
Support Vector Machine (SVM)	Full Text	Economy	0.75	0.86	0.8	28.0
Support Vector Machine (SVM)	Full Text	Other	0.88	0.7	0.78	33.0
Support Vector Machine (SVM)	Full Text	Politics	0.94	0.9	0.92	164.0
Support Vector Machine (SVM)	Full Text	Sports	0.91	0.78	0.84	41.0
Support Vector Machine (SVM)	Full Text	Macro Avg	0.84	0.83	0.83	400.0
Support Vector Machine (SVM)	Full Text	Weighted Avg	0.86	0.85	0.86	400.0
Support Vector Machine (SVM)	Stopwords Removed	Culture	0.74	0.91	0.82	67.0
Support Vector Machine (SVM)	Stopwords Removed	Disasters and Crimes	0.79	0.85	0.82	67.0
Support Vector Machine (SVM)	Stopwords Removed	Economy	0.77	0.82	0.79	28.0
Support Vector Machine (SVM)	Stopwords Removed	Other	0.87	0.61	0.71	33.0
Support Vector Machine (SVM)	Stopwords Removed	Politics	0.94	0.89	0.92	164.0
Support Vector Machine (SVM)	Stopwords Removed	Sports	0.89	0.83	0.86	41.0
Support Vector Machine (SVM)	Stopwords Removed	Macro Avg	0.83	0.82	0.82	400.0

Appendix 2. Performance of Supervised Machine Learning Classifiers Tested

Classifier	Preprocessing	Class	Precision	Recall	F1 Score	Support
Support Vector Machine (SVM)	Stopwords Removed	Weighted Avg	0.86	0.85	0.85	400.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Culture	0.69	0.9	0.78	67.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Disasters and Crimes	0.81	0.78	0.79	67.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Economy	0.9	0.68	0.78	28.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Other	1.0	0.58	0.73	33.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Politics	0.88	0.91	0.89	164.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Sports	0.92	0.88	0.9	41.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Macro Avg	0.87	0.79	0.81	400.0
Support Vector Machine (SVM)	Only Lead (First 75 words)	Weighted Avg	0.85	0.84	0.84	400.0