Project Report Phase 2

Martyna Wielgopolan, Aditya Kandula
June 1, 2025

Abstract

This project aims to develop a machine learning model capable of classifying news articles into predefined categories. Using a dataset of over 200,000 news headlines and descriptions, we explore multiple classification algorithms including Logistic Regression, Multinomial Naive Bayes, Support Vector Machines (SVM), and BERT. The goal is to evaluate their performance in terms of accuracy and generalization, with a specific focus on handling class imbalance and optimizing feature extraction through TF-IDF and deep embeddings.

---

1 Introduction

The increasing volume of online news has created a growing need for efficient automatic categorization. The task of news category prediction involves training a machine learning model to predict the most likely category a piece of news belongs to, based on its headline and description. In this project, we experiment with traditional classifiers and modern transformer-based models to determine the most effective approach for this task.

2 Overview of Used Algorithms

This project compares the performance of four distinct algorithms commonly used in text classification tasks. Each model represents a different approach to natural language processing, ranging from traditional statistical methods to deep learning.

- **Logistic Regression (LR):** A widely used linear model for classification, known for its simplicity, efficiency, and interpretability. It performs well with high-dimensional and sparse feature spaces, which makes it particularly suitable for TF-IDF-based text data.
- **Multinomial Naive Bayes (MNB):** A probabilistic model based on Bayes' theorem, often used as a strong baseline in text classification. It assumes conditional independence between features and works well when features represent word counts or term frequencies.
- **Support Vector Machine (SVM):** A margin-based linear classifier that attempts to find the hyperplane that best separates classes. SVMs are effective in high-dimensional spaces and are often used with TF-IDF representations in text classification problems.
- **BERT (Bidirectional Encoder Representations from Transformers):** A pre-trained deep language model developed by Google. Unlike traditional models, BERT captures the full context of a word using a transformer architecture, and can be fine-tuned for specific tasks such as classification with state-of-the-art performance.

These models span from fast and interpretable baselines to more complex and computationally intensive neural architectures. Each was evaluated on the same dataset split and preprocessed text data to ensure a fair comparison.

3 Dataset Description

The dataset used in this project is the *News Category Dataset* by Rishabh Misra, available on Kaggle. It contains approximately 120,000 news articles, each labeled with one of 41 topic categories. Each article includes a headline and a short description.
The dataset was randomly split into training and testing subsets as follows:

- **Training set:** 87% (approximately 104,400 samples)
- **Testing set:** 13% (approximately 15,600 samples)

Preprocessing Pipelines by Model

Different classification models required distinct text preprocessing strategies:

- **Logistic Regression and Multinomial Naive Bayes (MNB):** Text was fully normalized: lowercased, punctuation and digits removed, excessive whitespace reduced, and all non-ASCII characters stripped. This aggressive cleaning was designed to reduce feature sparsity for TF-IDF vectorization.
- **Support Vector Machine (SVM):** Preprocessing retained digits and focused on preserving structure useful for character-level TF-IDF features. Only punctuation and excess whitespace were removed.
- **BERT:** Minimal preprocessing was applied: only non-ASCII characters and excess whitespace were removed. This preserved the original structure of the input for the transformer-based tokenizer, which benefits from contextual richness and punctuation.

The preprocessing was tailored to each model's architecture. Classical ML models required feature compression and normalization, while transformer-based models leveraged raw textual patterns more effectively.

4 Implementation - Model Comparison and Evaluation

In this section, we describe the training process, tuning strategies, and observations for each of the four classification models used in this project: Logistic Regression, Multinomial Naive Bayes, Support Vector Machine (SVM), and BERT. All models were evaluated on the same stratified data split consisting of approximately 120,000 entries, with a train/test ratio of 87/13. However, each model used a distinct preprocessing pipeline optimized for its internal mechanisms.

Logistic Regression

Logistic Regression was used as a fast and effective linear baseline. We utilized TF-IDF features from both word- and character-level analyzers, with `max_features=20,000` each. Input text was lowercased and stripped of punctuation, numbers, and non-ASCII symbols.

- `C (inverse regularization strength)`: Optimal performance was observed around `C=1`. Smaller values underfit the data, while larger ones increased overfitting risk.
- `class_weight='balanced'` helped mitigate the effects of class imbalance by upweighting underrepresented categories.
- `solver='liblinear'` allowed efficient multiclass handling with L1 and L2 regularization.
- Other solvers such as `lbfgs` and `saga` were tested. While `lbfgs` performed similarly to `liblinear`, the `saga` solver showed promising theoretical potential (especially with large sparse datasets), but training was extremely slow and convergence was not achieved within reasonable time.

The model reached an accuracy of about **62%**. It was particularly strong on high-frequency classes but struggled with minority or semantically subtle categories.

Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) was chosen for its simplicity and efficiency. The same heavily normalized text as Logistic Regression was used.

- The smoothing parameter `alpha` played a crucial role. We tested several values, and found that `alpha=0.05` offered the best balance. Smaller values (e.g., 0.01) caused overfitting to frequent patterns, while larger values (e.g., 1.0) overly smoothed rare features.
- Although the MNB algorithm is theoretically better suited for raw count vectors, we empirically observed higher accuracy using TF-IDF vectors. This is likely due to the scale normalization reducing bias from very common n-grams.
- We experimented with different n-gram ranges and vector dimensions. Adding 2- or 3-grams improved performance slightly, but the gains diminished above 20,000 features.

Despite its simplicity, MNB achieved around **60% accuracy**, showing good recall across mid-frequency categories. However, it struggled with classes requiring deeper semantic understanding or longer context, which are inherently difficult for bag-of-words approaches.

Support Vector Machine (SVM)

For the Support Vector Machine model, we used the `LinearSVC` implementation with a TF-IDF feature representation combining both word- and character-level n-grams. Specifically:

- **TF-IDF features:** Word-level n-grams (1 to 2) with stop words removed and `max_features=15,000`, and character n-grams (3 to 5) with `max_features=5,000`.
- **Feature selection:** We applied `SelectKBest` with the chi-squared (`chi2`) test to retain the top 15,000 most informative features, reducing overfitting and speeding up training.
- **Classifier:** The final classifier was `LinearSVC` with `max_iter=200,000` to ensure convergence.

To improve performance, we experimented with:

- different n-gram ranges (e.g. word (1,3) and char (2,6)) — which increased training time but didn't significantly improve results;
- using or skipping `SelectKBest` — it helped reduce noise and improved F1-scores for low-resource classes;
- trying `class_weight='balanced'` — but it had minimal effect compared to Logistic Regression.

This setup led to an accuracy of about **62%**, with stronger performance than Naive Bayes on nuanced or longer text inputs, due to the richer character-level representation.

BERT (Transformer-Based Model)

For this project, the `bert-base-uncased` model was fine-tuned using the Hugging Face Transformers library. The input was a combination of the news *headline* and *short description*, with minimal cleaning (removing non-ASCII characters). Unlike older methods like TF-IDF, BERT learns the meaning of words based on their context.

- **Tokenizer:** Used `bert-base-uncased` tokenizer with special tokens and padding.
- **Input Length:** Limited to 128 tokens to keep training fast and efficient.
- **Batch Size:** 16 samples per batch to balance speed and memory use.
- **Learning Rate:** Set to 2×10^{-5}, a common choice for BERT fine-tuning.
- **Optimizer:** AdamW optimizer for better handling of BERT's weight updates.
- **Mixed Precision Training:** Used `torch.cuda.amp` to speed up training and save memory.
- **Epochs:** Trained for 3 passes (epochs) through the data.

The fine-tuned BERT model achieved about **74%** accuracy on the test set, which was much better than traditional machine learning models. It was especially good at handling difficult and less frequent news categories. To better understand how BERT separated the data, we used t-SNE and PCA to visualize the learned word embeddings.

Hyperparameter Sensitivity Analysis

We also checked how changing different settings (hyperparameters) would affect the model:

- **Batch Size:**
  - *Higher (32 or 64):* Training is faster but needs more GPU memory. Sometimes it might not generalize as well.
  - *Lower (8):* Training is slower but might give slightly better results because of more frequent updates.
- **Learning Rate:**
  - *Higher (5×10^{-5}):* Model learns faster but can become unstable and not finish properly.
  - *Lower (1×10^{-5}):* Model learns slower but is more stable. Needs more time to finish training.
- **Number of Epochs:**

  - *More Epochs (5 or more):* Can improve performance but also risks overfitting (model remembers training data too much).
  - *Fewer Epochs (1-2):* Faster training but model might not learn enough.
- **Maximum Token Length:**
  - *Longer (256 or 512):* More of the input text is kept, which helps if the news descriptions are long, but it also makes training slower.
  - *Shorter (64):* Faster and lighter training, but important parts of the text might be cut off.

Overall Observations

- **Preprocessing matters:** Each model benefits from tailored input cleaning. Heavy normalization helped LR/MNB, while BERT favored richer text.
- **Model complexity vs. performance:** Traditional models are fast and interpretable, but plateau in performance; BERT significantly raises accuracy at the cost of time and resources.
- **Hyperparameter tuning:** Small changes in regularization, feature size, and smoothing drastically impacted classical models. In contrast, BERT's performance was sensitive to batch size and learning rate.
- **Imbalanced classes:** All models struggled to some extent, but BERT handled them better without explicit rebalancing, likely due to its deep contextual understanding.

5 Conclusion

Through systematic preprocessing, feature extraction, and evaluation, we identified strengths and weaknesses in each algorithm. Logistic Regression remains a strong baseline for sparse TF-IDF features, while BERT significantly improves results using contextual embeddings.

BERT Classification Report (Best Model)

The BERT-based model trained using `train_bert(...)` with parameters: `epochs=3`, `batch_size=16`, `max_len=128`, and `lr=2e-5`, achieved the highest accuracy and overall performance across all categories. The table below presents the detailed classification report.

Category	Precision	Recall	F1-score	Support
ARTS	0.54	0.53	0.54	119
ARTS & CULTURE	0.51	0.55	0.53	120
BLACK VOICES	0.58	0.51	0.55	416
BUSINESS	0.68	0.56	0.62	606
COLLEGE	0.60	0.63	0.61	111
COMEDY	0.61	0.48	0.54	545
CRIME	0.59	0.76	0.67	293
CULTURE & ARTS	0.83	0.72	0.77	127
DIVORCE	0.86	0.90	0.88	401
EDUCATION	0.58	0.48	0.52	86
ENTERTAINMENT	0.74	0.86	0.79	1730
ENVIRONMENT	0.80	0.66	0.73	185
FIFTY	0.59	0.33	0.43	96
FOOD & DRINK	0.79	0.84	0.81	587
GOOD NEWS	0.37	0.52	0.43	143
GREEN	0.52	0.51	0.51	234
HEALTHY LIVING	0.64	0.64	0.64	654
HOME & LIVING	0.89	0.78	0.83	325
IMPACT	0.48	0.47	0.48	317
LATINO VOICES	0.47	0.48	0.47	108
MEDIA	0.60	0.54	0.57	270
MONEY	0.68	0.66	0.67	190
PARENTING	0.88	0.76	0.82	895
PARENTS	0.79	0.54	0.64	385
POLITICS	0.84	0.84	0.84	3518
QUEER VOICES	0.82	0.74	0.78	619
RELIGION	0.77	0.58	0.66	232
SCIENCE	0.62	0.51	0.56	181
SPORTS	0.70	0.83	0.76	493
STYLE	0.72	0.62	0.67	268
STYLE & BEAUTY	0.90	0.90	0.90	907
TASTE	0.45	0.59	0.51	217
TECH	0.52	0.67	0.59	174
THE WORLDPOST	0.59	0.64	0.61	336
TRAVEL	0.83	0.90	0.86	996
U.S. NEWS	0.47	0.30	0.36	115
WEDDINGS	0.88	0.92	0.90	436
WEIRD NEWS	0.46	0.41	0.43	257
WELLNESS	0.84	0.92	0.88	1699
WOMEN	0.59	0.41	0.49	342
WORLD NEWS	0.48	0.49	0.48	267
WORLDPOST	0.66	0.76	0.71	234

Table 1: Classification report for BERT-based model - best performing among all tested ones - accuracy (74%).

Logistic Regression Classification Report (Best Accuracy)

The Logistic Regression model, trained with a TF-IDF character + word n-gram representation, achieved the highest test accuracy of **62%**. The classification metrics for each category are shown below.

Category	Precision	Recall	F1-score	Support
ARTS	0.24	0.35	0.29	119
ARTS & CULTURE	0.25	0.36	0.29	120
BLACK VOICES	0.49	0.48	0.49	416
BUSINESS	0.51	0.50	0.50	606
COLLEGE	0.42	0.63	0.50	111
COMEDY	0.49	0.45	0.47	545
CRIME	0.45	0.63	0.53	293
CULTURE & ARTS	0.48	0.61	0.54	127
DIVORCE	0.80	0.79	0.79	401
EDUCATION	0.29	0.49	0.36	86
ENTERTAINMENT	0.74	0.62	0.67	1730
ENVIRONMENT	0.48	0.56	0.52	185
FIFTY	0.18	0.25	0.21	96
FOOD & DRINK	0.66	0.69	0.67	587
GOOD NEWS	0.25	0.37	0.30	143
GREEN	0.36	0.43	0.39	234
HEALTHY LIVING	0.52	0.43	0.47	654
HOME & LIVING	0.67	0.76	0.71	325
IMPACT	0.36	0.40	0.38	317
LATINO VOICES	0.35	0.41	0.38	108
MEDIA	0.43	0.57	0.49	270
MONEY	0.43	0.59	0.50	190
PARENTING	0.69	0.66	0.68	895
PARENTS	0.52	0.49	0.51	385
POLITICS	0.86	0.68	0.76	3518
QUEER VOICES	0.78	0.70	0.74	619
RELIGION	0.50	0.56	0.53	232
SCIENCE	0.39	0.50	0.44	181
SPORTS	0.63	0.68	0.65	493
STYLE	0.42	0.56	0.48	268
STYLE & BEAUTY	0.87	0.80	0.83	907
TASTE	0.26	0.31	0.28	217
TECH	0.34	0.48	0.40	174
THE WORLDPOST	0.48	0.49	0.49	336
TRAVEL	0.78	0.72	0.74	996
U.S. NEWS	0.16	0.19	0.17	115
WEDDINGS	0.80	0.88	0.84	436
WEIRD NEWS	0.31	0.38	0.34	257
WELLNESS	0.78	0.79	0.78	1699
WOMEN	0.42	0.54	0.47	342
WORLD NEWS	0.38	0.37	0.38	267
WORLDPOST	0.43	0.52	0.47	234

Table 2: Classification report for Logistic Regression model with highest accuracy (62%).

Multinomial Naive Bayes Classification Report

The Multinomial Naive Bayes model trained with TF-IDF character and word n-grams achieved a test accuracy of **60%**. The classification report per category is shown below.

Category	Precision	Recall	F1-score	Support
ARTS	0.39	0.27	0.32	119
ARTS & CULTURE	0.30	0.17	0.22	120
BLACK VOICES	0.48	0.38	0.42	416
BUSINESS	0.49	0.54	0.51	606
COLLEGE	0.45	0.52	0.48	111
COMEDY	0.45	0.45	0.45	545
CRIME	0.44	0.67	0.53	293
CULTURE & ARTS	0.43	0.53	0.47	127
DIVORCE	0.68	0.78	0.73	401
EDUCATION	0.33	0.34	0.33	86
ENTERTAINMENT	0.65	0.68	0.67	1730
ENVIRONMENT	0.51	0.43	0.47	185
FIFTY	0.38	0.14	0.20	96
FOOD & DRINK	0.58	0.75	0.65	587
GOOD NEWS	0.34	0.34	0.34	143
GREEN	0.33	0.34	0.34	234
HEALTHY LIVING	0.46	0.31	0.37	654
HOME & LIVING	0.64	0.65	0.65	325
IMPACT	0.39	0.38	0.39	317
LATINO VOICES	0.41	0.19	0.26	108
MEDIA	0.45	0.44	0.44	270
MONEY	0.45	0.45	0.45	190
PARENTING	0.53	0.64	0.58	895
PARENTS	0.49	0.39	0.43	385
POLITICS	0.79	0.76	0.78	3518
QUEER VOICES	0.70	0.59	0.64	619
RELIGION	0.53	0.44	0.48	232
SCIENCE	0.55	0.44	0.49	181
SPORTS	0.65	0.63	0.64	493
STYLE	0.49	0.43	0.46	268
STYLE & BEAUTY	0.79	0.73	0.76	907
TASTE	0.29	0.22	0.25	217
TECH	0.43	0.40	0.42	174
THE WORLDPOST	0.41	0.46	0.43	336
TRAVEL	0.67	0.76	0.71	996
U.S. NEWS	0.14	0.14	0.14	115
WEDDINGS	0.77	0.78	0.78	436
WEIRD NEWS	0.38	0.32	0.35	257
WELLNESS	0.64	0.76	0.70	1699
WOMEN	0.47	0.47	0.47	342
WORLD NEWS	0.34	0.31	0.33	267
WORLDPOST	0.50	0.30	0.38	234

Table 3: Classification report for Multinomial Naive Bayes model (accuracy: 60%).

Linear SVM Classification Report

The Linear Support Vector Machine (SVM) model achieved a test accuracy of **62%**. Below is the detailed classification report per category.

Category	Precision	Recall	F1-score	Support
ARTS	0.26	0.27	0.27	119
ARTS & CULTURE	0.23	0.25	0.24	120
BLACK VOICES	0.43	0.47	0.45	416
BUSINESS	0.49	0.49	0.49	606
COLLEGE	0.44	0.59	0.50	111
COMEDY	0.45	0.46	0.46	545
CRIME	0.48	0.61	0.54	293
CULTURE & ARTS	0.51	0.54	0.52	127
DIVORCE	0.77	0.81	0.79	401
EDUCATION	0.28	0.34	0.31	86
ENTERTAINMENT	0.73	0.65	0.68	1730
ENVIRONMENT	0.51	0.46	0.48	185
FIFTY	0.19	0.21	0.20	96
FOOD & DRINK	0.64	0.72	0.68	587
GOOD NEWS	0.26	0.34	0.30	143
GREEN	0.31	0.36	0.34	234
HEALTHY LIVING	0.51	0.44	0.47	654
HOME & LIVING	0.70	0.75	0.73	325
IMPACT	0.32	0.38	0.35	317
LATINO VOICES	0.32	0.35	0.33	108
MEDIA	0.40	0.52	0.45	270
MONEY	0.49	0.56	0.52	190
PARENTING	0.68	0.66	0.67	895
PARENTS	0.47	0.44	0.46	385
POLITICS	0.85	0.71	0.77	3518
QUEER VOICES	0.73	0.73	0.73	619
RELIGION	0.48	0.53	0.50	232
SCIENCE	0.43	0.46	0.44	181
SPORTS	0.64	0.68	0.66	493
STYLE	0.45	0.53	0.49	268
STYLE & BEAUTY	0.86	0.83	0.84	907
TASTE	0.26	0.28	0.27	217
TECH	0.37	0.41	0.39	174
THE WORLDPOST	0.41	0.47	0.44	336
TRAVEL	0.76	0.74	0.75	996
U.S. NEWS	0.16	0.16	0.16	115
WEDDINGS	0.80	0.87	0.84	436
WEIRD NEWS	0.29	0.31	0.30	257
WELLNESS	0.79	0.79	0.79	1699
WOMEN	0.40	0.50	0.44	342
WORLD NEWS	0.34	0.35	0.34	267
WORLDPOST	0.45	0.49	0.47	234

Table 4: Classification report for Linear SVM model (accuracy: 62%).