NLP Assignment 1 Group 6

Cristian Alin Aron Enzo Castelo Branco Biondi (S6579434)
Bogdan-Gabriel Ivancu (S5160901)

February 2026

1 Introduction

Natural Language Processing (NLP) is an area of machine learning that allows computers to work with human language. In this assignment, a classical machine learning pipeline to text classification was implemented, using the AG News dataset to label news articles into four categories: World, Sports, Business, and Sci/Tech. The goal was to establish a solid baseline using traditional methods like Logistic Regression and SVM, which can be used as reference point when evaluating future improvements to the pipeline.

2 Dataset and Split Details

The AG News dataset consists of news articles labeled into four categories: (1) World, (2) Sports, (3) Business, and (4) Sci/Tech. Each instance contains the article title and a short description.

The dataset provides a predefined split into a training set with 120,000 instances and a test set with 7,600 instances. From the original training set, a validation (development) set was created using a 90/10 split with a fixed random seed of 36, resulting in 108,000 training instances and 12,000 validation instances.

3 Preprocessing

At an initial inspection, the dataset appeared clean, with no missing or null values. However, a closer examination of the text revealed minor inconsistencies in the article descriptions, such as the presence of stray characters (e.g., "/" appearing between words). These issues were addressed during preprocessing to ensure more consistent textual input.

After text normalization, tokenization and feature extraction were performed using `TfidfVectorizer` from `scikit-learn`, with a maximum vocabulary size of 10,000 features.

4 Modeling

Two linear classifiers were trained for the task: Logistic Regression and a linear Support Vector Machine (SVM). For both models, a small grid search was performed on the regularization parameter C, which controls the trade-off between model complexity and fitting the training data.

The results indicate that Logistic Regression performs best with $C = 1$, achieving an accuracy of 0.9077 and a Macro-F1 score of 0.9081 on the development set. For the linear SVM, stronger regularization ($C = 0.1$) yielded the best results, resulting in a development accuracy of 0.9099 and a Macro-F1 score of 0.9103.

5 Results

After model selection on the development set, the best-performing configuration was evaluated once on the held-out test set. The final model achieved an accuracy of 0.9147 and a Macro-F1 score of 0.9145, indicating strong and balanced performance across all four classes.

Figure 1 shows the confusion matrix for the test set. Overall, the model performs well across all categories, with most predictions concentrated along the main diagonal. The largest confusion patterns occur between the *Business* and *Sci/Tech* classes, suggesting semantic overlap between economic and technology-related news articles.

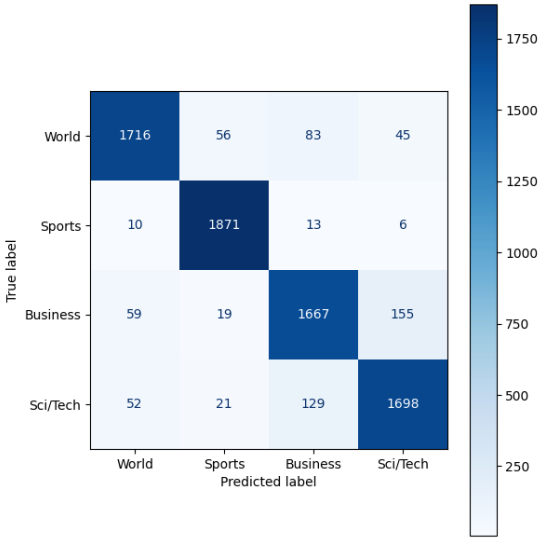

Figure 1: Confusion Matrix (Test Set)

6 Error Analysis

To further understand the errors of the model, the misclassified examples were analyzed, and 3 reoccurring patterns were recognized.

1. **Business vs. Sci/Tech overlap:** The most common form of error, usually happens to articles from technology companies and products that also involve finance terms.
2. **World vs. Business overlap:** Most of the mistakes in this category revolve around articles involving economic policies and trade agreements.
3. **Sports-related non-sports context:** Articles that are not sports related but involve athletes or sports organizations often get classified as sports.
4. **Label noise in the dataset:** A small number of misclassified examples appear to contain genuine inconsistencies in the dataset labels. These cases are unresolvable by any word-level model and place a ceiling on achievable accuracy.

Table 1: Representative misclassified test examples grouped by error type.

Text (truncated)	**True**	**Pred**	**Type**
Text (truncated)	**True**	**Pred**	**Type**
Google tests book search — new tool matches queries with book content	Business	Sci/Tech	1
Global server sales up in second quarter — rose 7.7% to $11.55 billion	Sci/Tech	Business	1
Motorola acquires wireless mesh vendor — will acquire MeshNetworks, developer of 802.11 mesh	Business	Sci/Tech	1
Nortel reports loss in Q3, sees pickup in Q4 — reported a loss for its third quarter	Sci/Tech	Business	1
Google shares just may be winners — Wall Street forced Google to lower its IPO share price	Sci/Tech	Business	1
McAfee profit growth offsets AOL questions — shares up on stronger-than-expected earnings	Sci/Tech	Business	1
Ryanair, EasyJet expand after Volare bankruptcy — both announced plans to expand in Italy	World	Business	2
Cuba developing trade with US companies despite blockade	Business	World	2
Stocks dip on consumer income report — unsettling data set off profit-taking on Wall St	World	Business	2
Test of US missile defence shield fails — interceptor launch attempt failed	Business	World	2

Deal nears on Iraq debt write-off — Paris Club to forgive 80% of Iraq's debt at G20	Business	World	2
Red Sox seek approval to expand Fenway capacity	Business	Sports	3
The division is almost in the Yankees' hands — after a doubleheader sweep	World	Sports	3
Sox make off-season pitch — considering Fenway hall of fame to keep cash flowing	Business	Sports	3
With boxing show, Fox makes quick move — premiere of 'The Next Great Champ'	Business	Sports	3
Buying hockey gear with Cam Neely — labor dispute may sideline professional hockey	Business	Sports	3
It's tax-loss season, y'all — got winners? got losers? time to save a tax dollar	Sports	World	4
Report: Palestinians, Israel back peace plan — agreed in principle to end-conflict proposals	Business	Sci/Tech	4
Ohio St. Buckeyes — Ohio State hopes to avoid back-to-back defeats under Tressel	Business	Sci/Tech	4
Thousands march in London to protest Iraq war — marched through central London	Business	Sci/Tech	4

Overall, the observed errors are consistent with the limitations of word-level TF-IDF features. Most of the confusion between *Business* and *Sci/Tech* comes from articles about technology companies and products that mix technical and financial vocabulary. Similarly, the overlap between *World* and *Business* reflects how economic policies and international trade topics are often closely connected, making them difficult to separate based only on words. The sports-related errors further show that the model relies heavily on individual keywords, struggling with articles that mention teams or athletes even when sports are not the main subject. These patterns suggest that richer representations capturing deeper context would be needed to reduce those errors.

7 Reproducibility Notes

All experiments were run on Python 3.12 The full source code and environment file are available at:

$$\texttt{https://github.com/BogdanI40/NLP}_{Group}6$$

The repository has the following structure:

```
Assignment1/
  README.md            <- run instructions and results summary
  requirements.txt     <- exact package versions
  assignment1.py    <- full pipeline (data -> features -> model -> eval)
  report/
    NLP_Group6.pdf  <- report
```

1. Install dependencies.

```
pip install -r requirements.txt
```

2. Run.

```
python assignment1.py
```

The code automatically downloads the AG News dataset from HuggingFace, performs preprocessing, trains both models, and outputs all metrics and figures. Expected runtime: approximately 3–5 minutes on a standard CPU.

Key reproducibility safeguards.

Table 2: Reproducibility safeguards.

Safeguard	**Detail**
Random seed	`SEED = 36` used in `train_test_split` and `LinearSVC(random_state=SEED)`
Vectoriser fit	`TfidfVectorizer` fitted on train only; dev and test are transformed only
Hyperparameter search	5-fold CV on train set only; dev used solely for final model selection
Test set policy	Test evaluated exactly once, after model selected on dev set
Dataset source	HuggingFace: `sh0416/ag_news` (official AG News predefined split)