NLP Assignment 2023

AG News Data

Group 10
Adrià Termes
Ilias Trakas
Federico Colombo
Nathaniel Thomas-Copeland
Diego Sánchez

Agenda

1. Choosing a General Model to Tackle Individual Category Ones
2. Evaluating the Performance of Individual Models
3. Quick Overview of Coefficients Importance
4. Conclusions and Next Steps

Deciding the individual category models: explainability wins and accuracy tradeoff is softened thanks to hyperparameterisation

- Usage of CountVectorizer() and TFIDF vectorizer, reduces stop words appearance.

- Considerable levels of precision and recall, especially for sports, with an overall accuracy of 90%.

Naive Bayes Multinomial

Low explainability in feature importance because of the naive assumption of independence in the model: we disregard it.

	precision	recall	f1-score	support
science	0.86	0.80	0.83	2537
sports	0.86	0.98	0.92	2458
world	0.88	0.85	0.87	2509
business	0.86	0.82	0.84	2496
accuracy			0.86	10000
macro avg	0.86	0.87	0.86	10000
weighted avg	0.86	0.86	0.86	10000

SGD Classifier

High explainability, low accuracy (86%). Playing with GridSearchCV, trigrams and alpha increase it.

Best Parameters: {'clf__alpha': 1e-06, 'vect__ngram_range': (1, 3)}

	precision	recall	f1-score	support
science	0.89	0.89	0.89	2537
sports	0.95	0.98	0.97	2458
world	0.91	0.90	0.91	2509
business	0.89	0.87	0.88	2496
accuracy			0.91	10000
macro avg	0.91	0.91	0.91	10000
weighted avg	0.91	0.91	0.91	10000

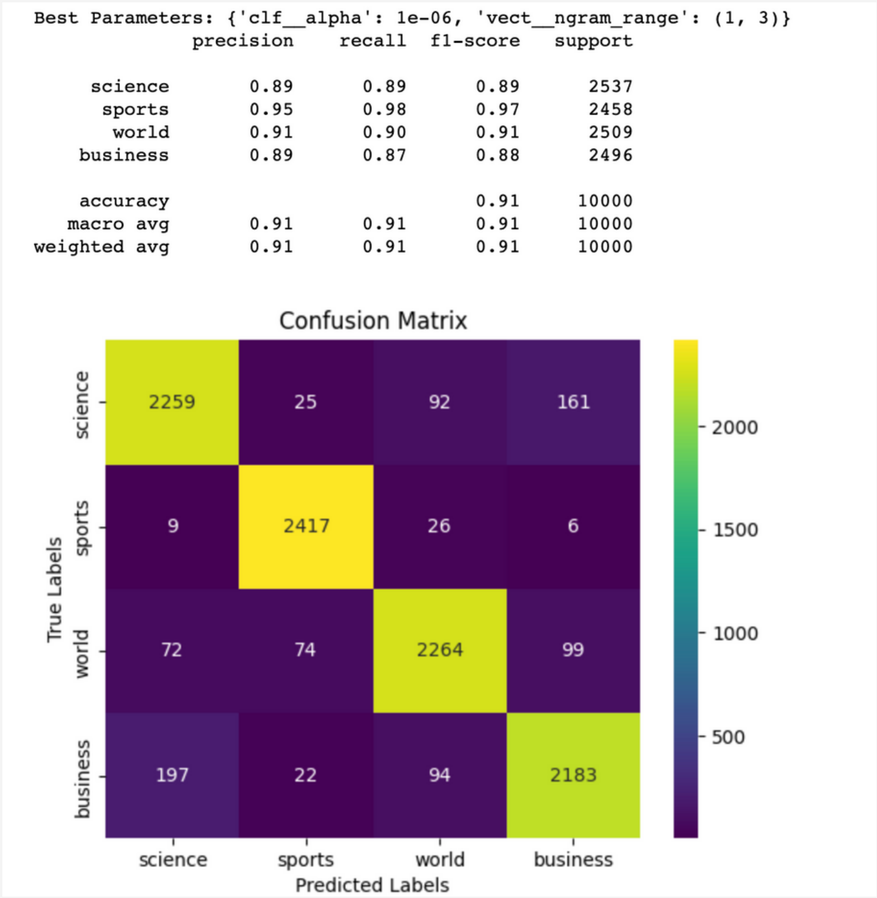
Individual category SGD: Similar metrics, slight decrease in recall but better precision metrics. Sports, easily spotted.

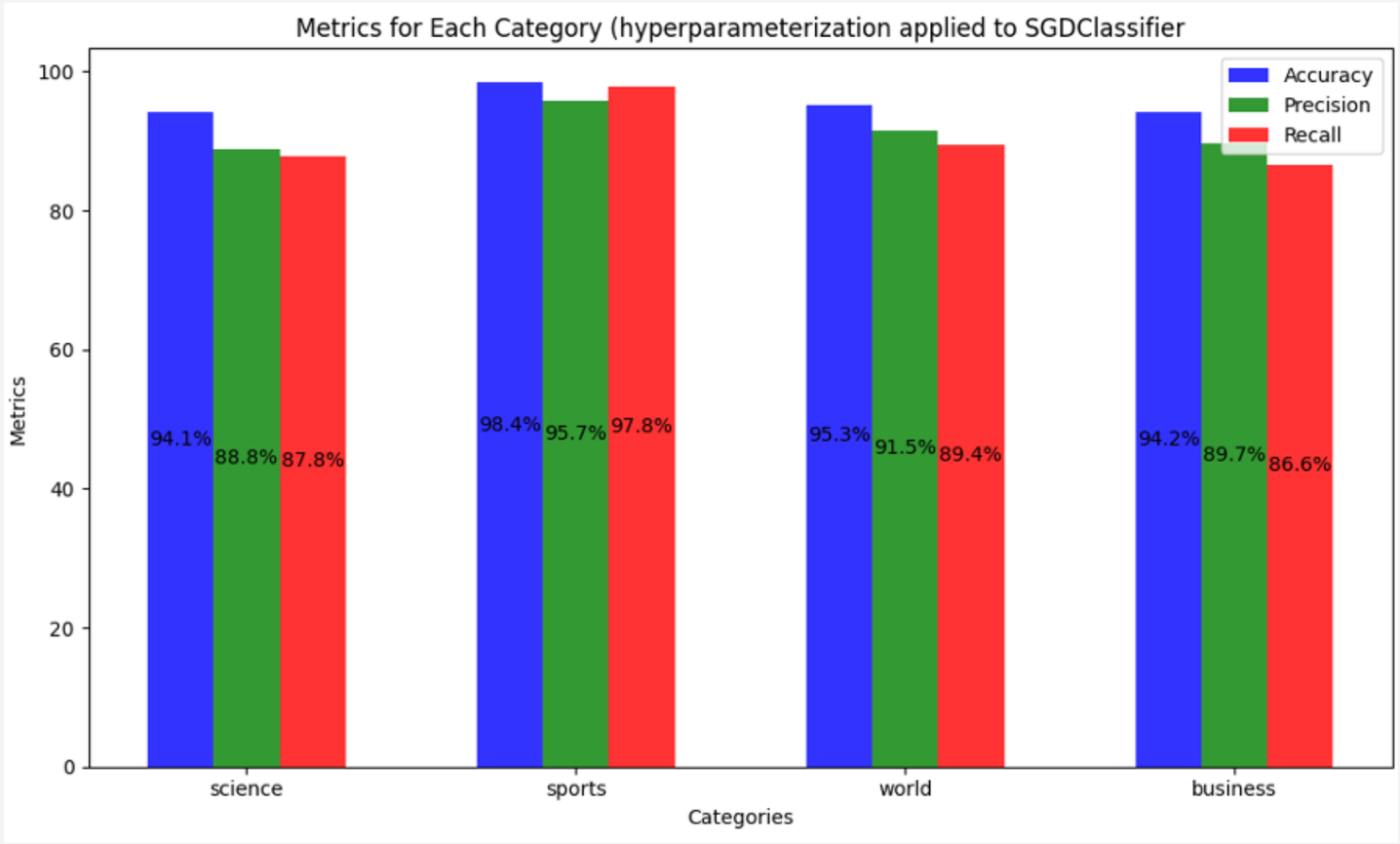
- GridSearchCV, with CV of 5.
- Trigrams are selected in all models.

```
Best Parameters: {'clf__alpha': 1e-06, 'vect__ngram_range': (1, 3)}
              precision    recall  f1-score   support

     science       0.89      0.89      0.89      2537
      sports       0.95      0.98      0.97      2458
       world       0.91      0.90      0.91      2509
    business       0.89      0.87      0.88      2496

    accuracy                           0.91     10000
   macro avg       0.91      0.91      0.91     10000
weighted avg       0.91      0.91      0.91     10000
```





Two different use cases for bigrams and trigrams, with suspicious structures pointing for or against a category being common

	Unigrams	Bigrams	Trigrams
	✓ space, software, nasa, linux...	✓ *washingtonpost com, ft news, open source, video game...	✓ *reuters intl business, lt gt san, 36 400 million, oracle corp 36...
	✓ coach, football, cup, team, olympic, season, ap...	✓ kobe bryant, manchester utd., world cup, us open...	✗ *greece michael phelps, game to win, elite sports the...
	✓ *afp, iraq, president, peace, elections, terrorism, baghdad...	✓ *canadian press, ap tokyo, north korea, united nations...	✓ *new york stocks, ap president bush, brasilia brazil reuters...
	✗ ap, afp, team, game, linux, nasa, washingtonpost, reuters...	✗ intel corp, reuters microsoft, manchester united, ap federal...	✗ e

*Denotes Suspicious values on the alternative side

Conclusions

Therefore, Spacy, multioutput models and CNN's should be also explored.

Semantic identification

Frequent words are conjunctions and prepositions. To improve semantic identification, deeper meanings or n-grams should be considered.

Models performance

Similar metrics, with sports consistently outperforming science and business in recall and precision. Individual models do not add up very much in metrics.

Semantics Categorisation

Although metrics are good, explainability is somehow clunky; with constant words that are strangely classified in coefficient terms, specially in trigrams.