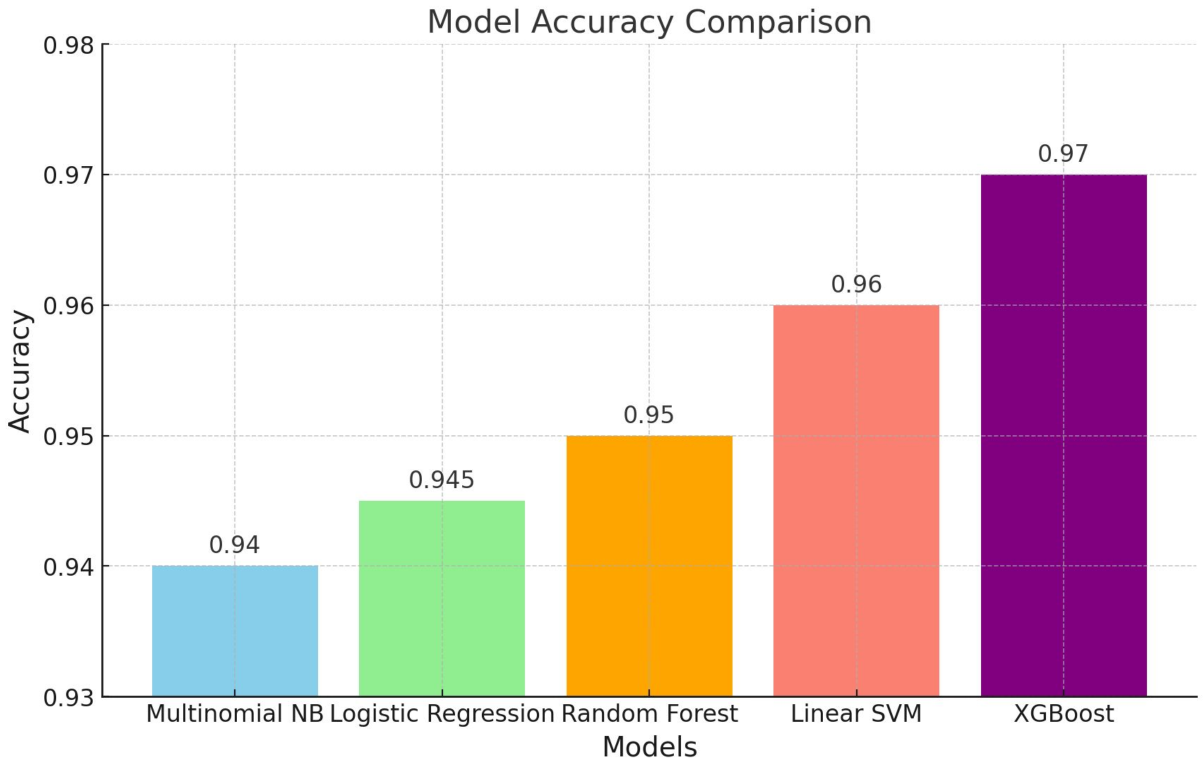
Group 1: NLP News Classifier

Sofia Zogkza, Khushboo Shrivastava, Inna Ivanova, Ramzi Ben Meftah

Executive summary

- Final result: 97% accuracy achieved
- Models used
  - Multinomial Naive Bayes,
  - Logistic Regression,
  - Random Forest,
  - Linear SVM,
  - XGBoost
- We used data preprocessing, TF-IDF, embeddings



Methods (preprocessing)

- How you approached the dataset, cleaning, preprocessing
  - tokenization
  - stop words removal
  - punctuation removal
  - lemmatization

- Techniques we tried
  - TF-IDF
  - Bag of Words
  - Word Embeddings

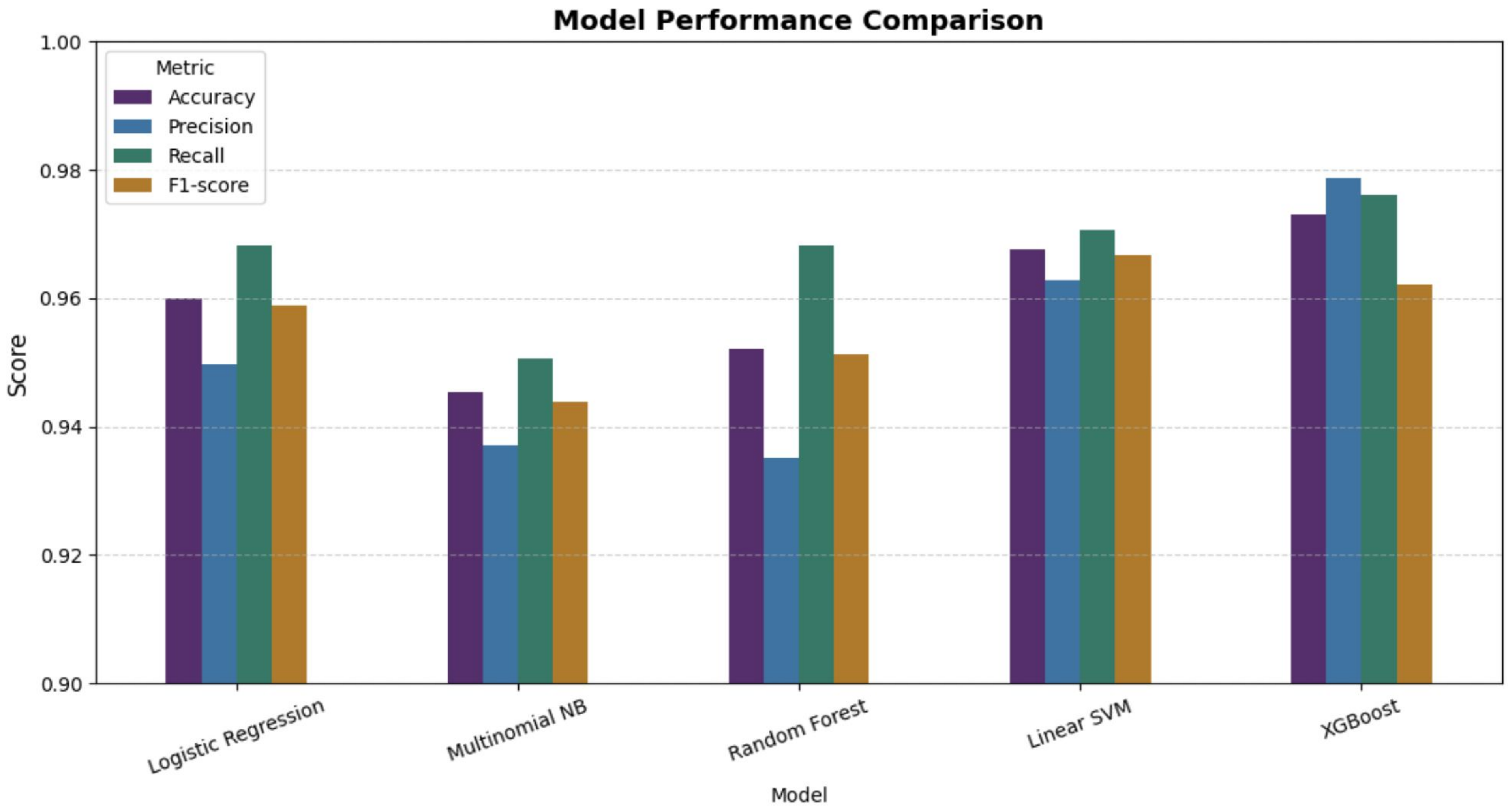
Methods (embedding)

- We used the pre-processed data for embeddings
- We used hybrid embedding with TF-IDF
- We tried both Word2Vec and GloVe
- Our final embedding technique was GloVe

Models – 1

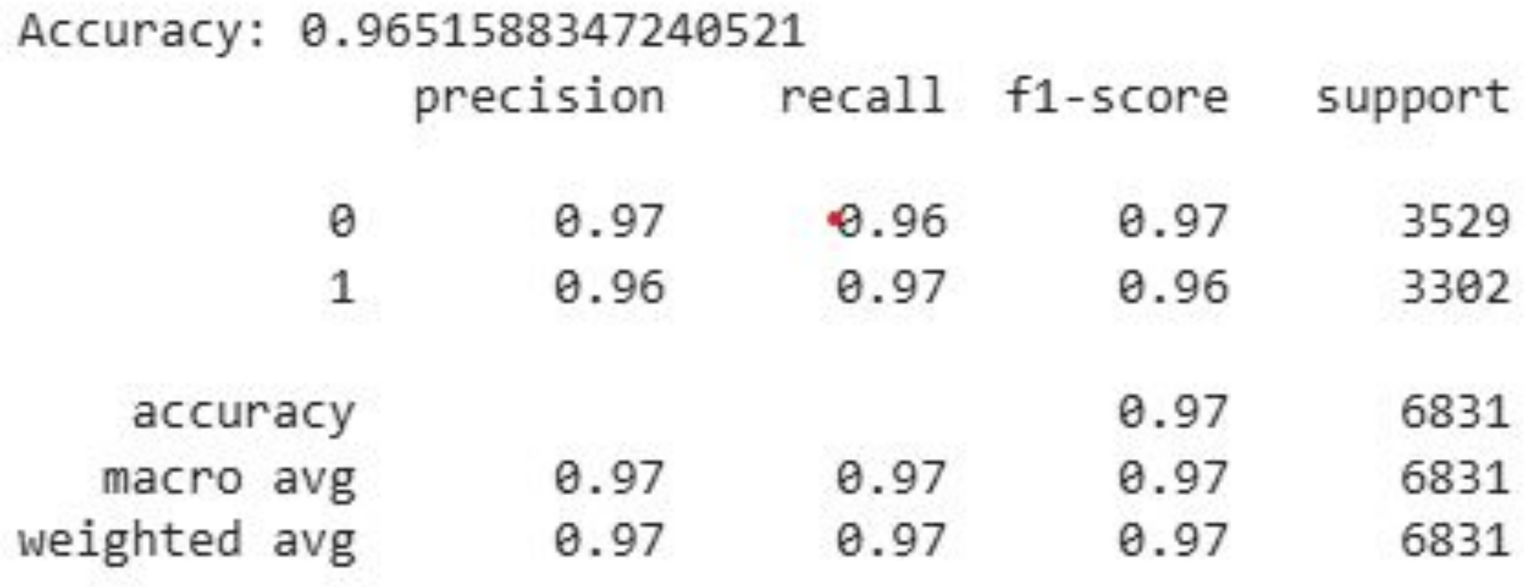
- Our experiments with several machine learning models
  - Multinomial Naive Bayes - baseline performance
  - Logistic Regression - gave better result than baseline
  - Random Forest - slow, didn’t bring improvement
  - Linear SVM - performance was very good
  - XGBoost - achieved the best results, slightly outperforming Linear SVM

Models - 2


Model Performance Comparison
Metric
Accuracy
Precision
Recall
F1-score
Score
1.00
0.98
0.96
0.94
0.92
0.90
Logistic Regression
Multinomial NB
Random Forest
Linear SVM
XGBoost
Model

Final Result with XGBoost

Model Evaluation Metrics



```
Accuracy: 0.9651588347240521
              precision    recall  f1-score   support

           0       0.97      0.96      0.97      3529
           1       0.96      0.97      0.96      3302

    accuracy                           0.97      6831
   macro avg       0.97      0.97      0.97      6831
weighted avg       0.97      0.97      0.97      6831
```

Takeaways

- Recap / conclusions
  - Preprocessing quality directly impacts performance
  - ML deliver good results, without the complexity of NN
- Challenges
  - Preprocessing
  - Embeddings
- Key learnings
  - XGBoost achieved highest accuracy
  - Linear SVM takes the second place with good generalization(after good preprocessing)

Thank you