GROUP 1

NLP model

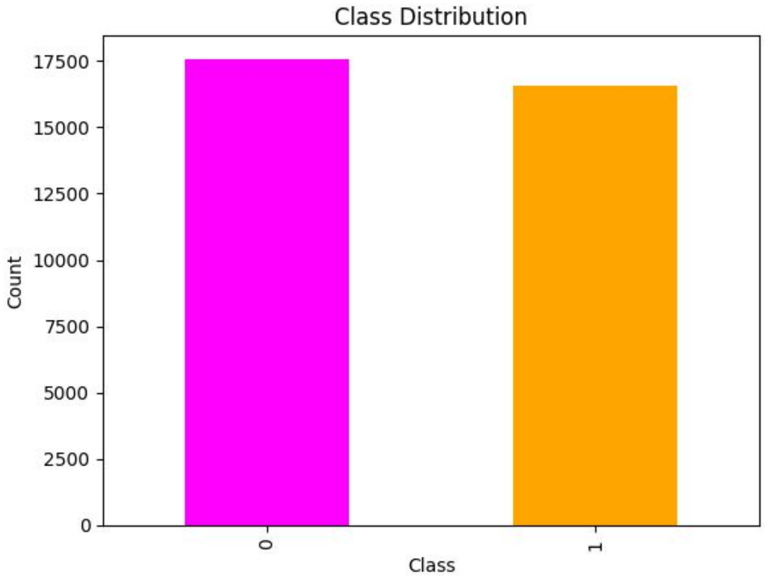
Freddy Roldán Saiqa Mehdi Marina Castillo

1. Executive summary

- Final result: **91,70%** accuracy
- Model used:
  - **Naive Bayes** + **CountVectorizer**
- Tried:
  - Naive Bayes + TF-IDF / + CountVectorizer
  - Random Forest + TF-IDF / + CountVectorizer

2. Data Preprocessing and Feature Engineering

- Data exploration
- Lemmatization
- Special Character Removal
- StopWords Removal
- Tokenization
- WordNet

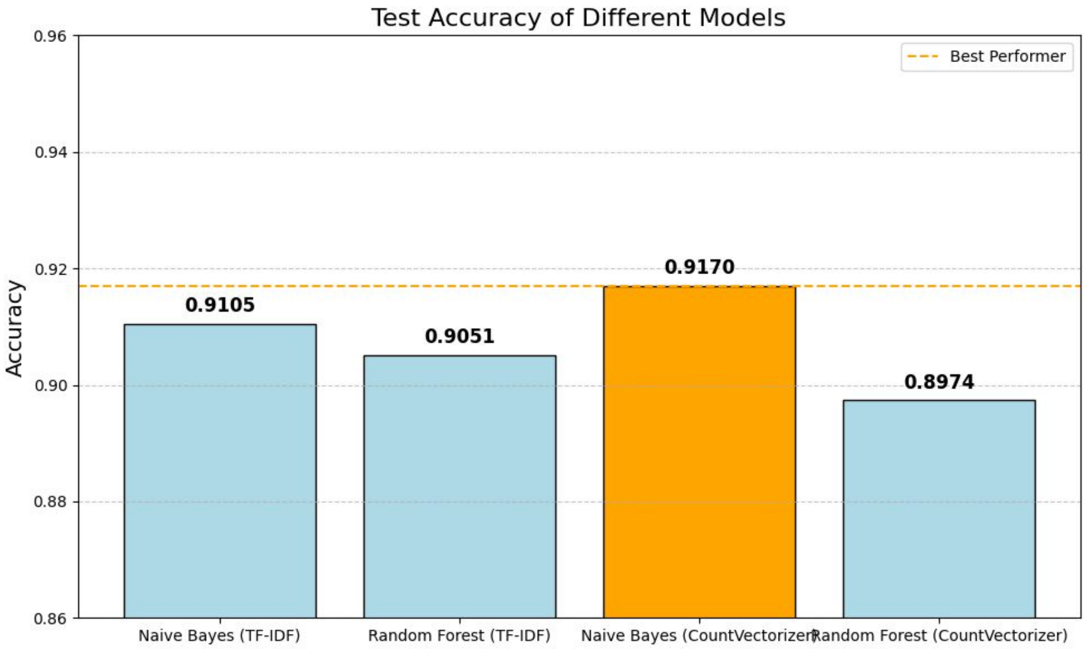


2. Data Preprocessing and Feature Engineering

- TF-IDF and CountVectorizer (Faster and Accurate)
- Sentiment Analysis (TextBlob)

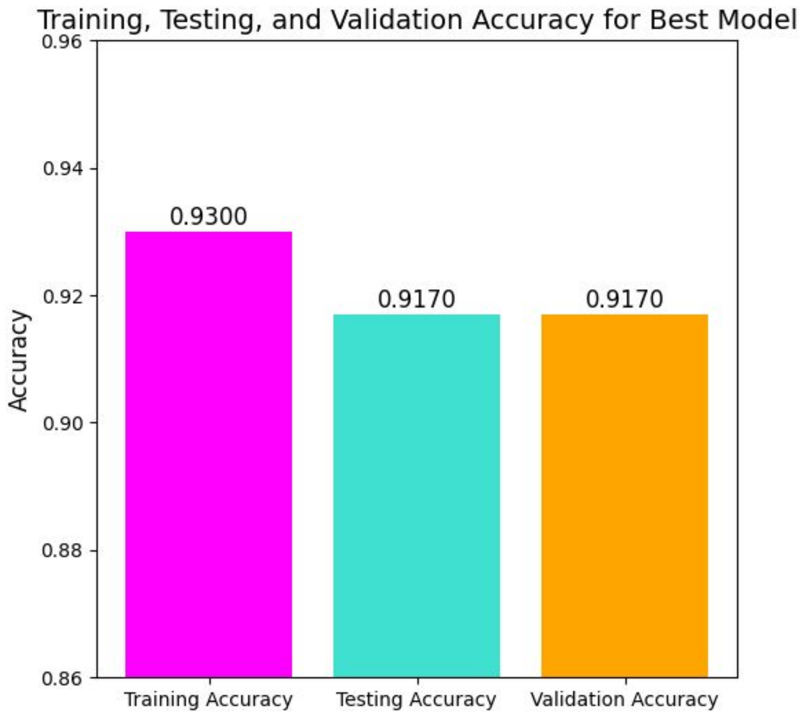
3. Modeling and Training

- **Models** used: Naive Bayes and Random Forest



4. Results of Naive Bayes + CountVectorizer

- Cross-Validation Accuracy: **0.9170**
- Cross-Validation F1 Score: **0.9166**
- Insights:
  - Effective Text **Classification**
  - Strong **Generalization** Capability
  - **Consistent Performance** Across Cross-Validation Folds
- Comparison: **Slightly lower** than training accuracy (0.9300)
  - Good **generalization**
  - Slightly **overfitting**



5. Takeaways

- Recap / conclusions
- Challenges
  - Compatibility
  - Negative Values

- Key learnings
  - Time management

- Steps to improve project:
  - Hyperparameter Tuning
  - Use Pre Trained Embeddings
  - More Complex Models

Thank you.

Questions?