LARA Team

PROJECT 2: NLP CHALLENGE—FAKE NEWS CLASSIFICATION

MEMBERS:

Alrumaysaa Alghamdi

Layla Alsulaimani

Razan Alkhamisi

Contents

Project Introduction

Dataset

Preprosses

Classical ML

Results

Word2Vec-based classifier

Training Results

Other Results

Overall Results

Demo

Project Introduction

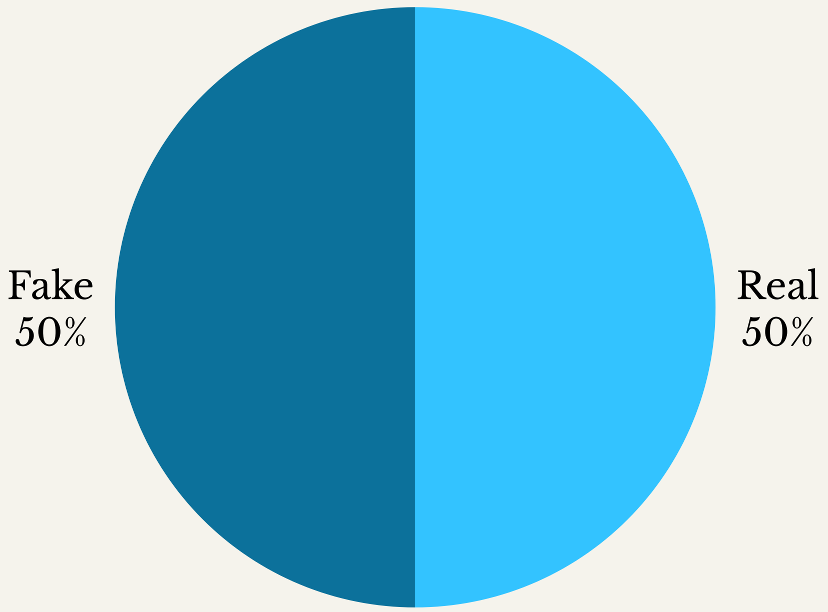
Project Objective:
Build a classifier to detect whether a news article is real or fake.

Approach:

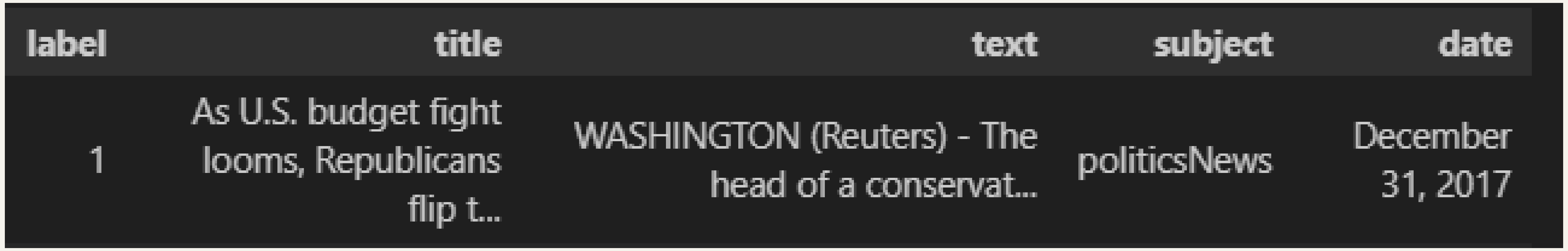
- Develop a classical NLP model.
- Develop a Word2Vec-based classifier.
- Compare between models performance.

Dataset

- **Total Records**: 39,942
- **Label Distribution:**
    - Real News (1): 19,999
    - Fake News (0): 19,943



- **Columns:**
    - **label:** (0 = Fake, 1 = Real)
    - **title:** News headline
    - **text**: Full article content
    - **subject**: News category
    - **date:** Publication date



label	title	text	subject	date
1	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017

Preprocessing Steps

1 **Convert to lowercase:**
To standardizes text.

2 **Remove Special Characters & Numbers:**
To keeps only letters.

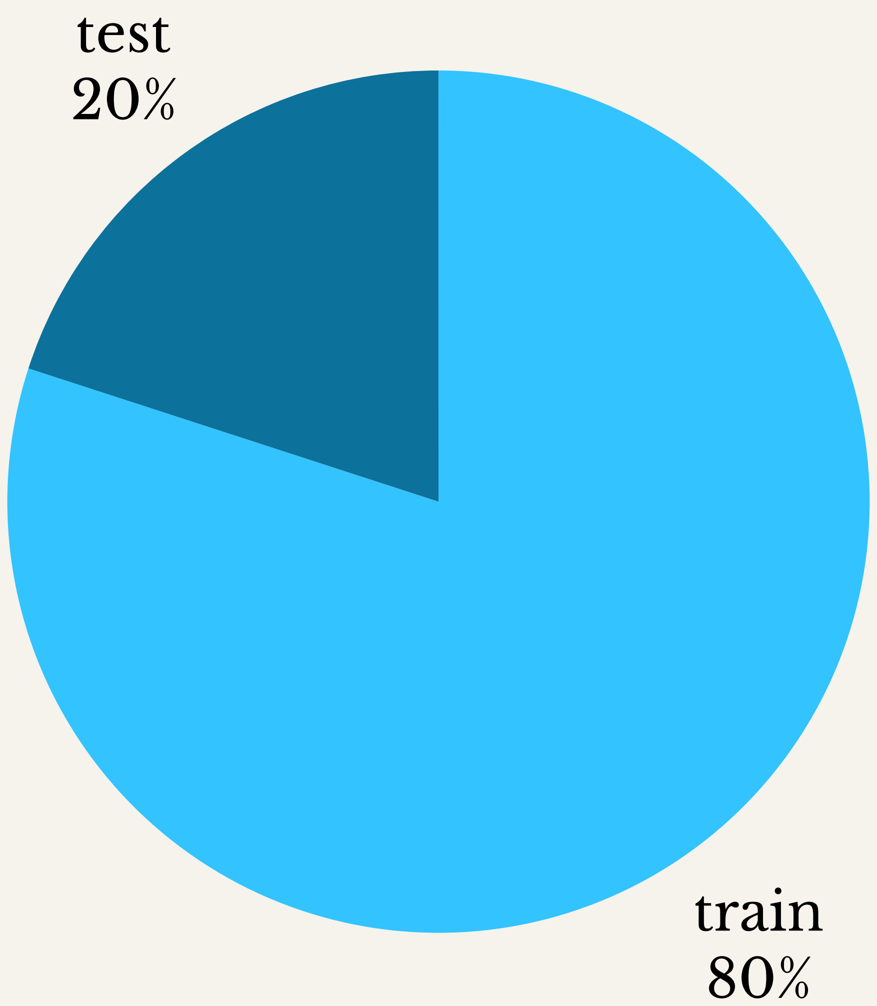
3 **Tokenization & Stopword Removal:**
To breaks text into words & removes common words.
`(stopwords.words('english'))`

4 **Lemmatization:**
Reduces words to their base form.
`(WordNetLemmatizer())`

5 **Feature Extraction (TF-IDF):**
Converts text into numerical representation.

Feature Extraction & Data Splitting

1 Splitting The Data



2 TF-IDF Vectorization:

- Uses unigrams & bigrams (ngram_range=(1,2)) for better feature representation.
- Filters terms appearing in less than 2% or more than 95% of documents (max_df=0.95, min_df=0.02).
- Converts raw text into numerical vectors for model training.

Classical ML

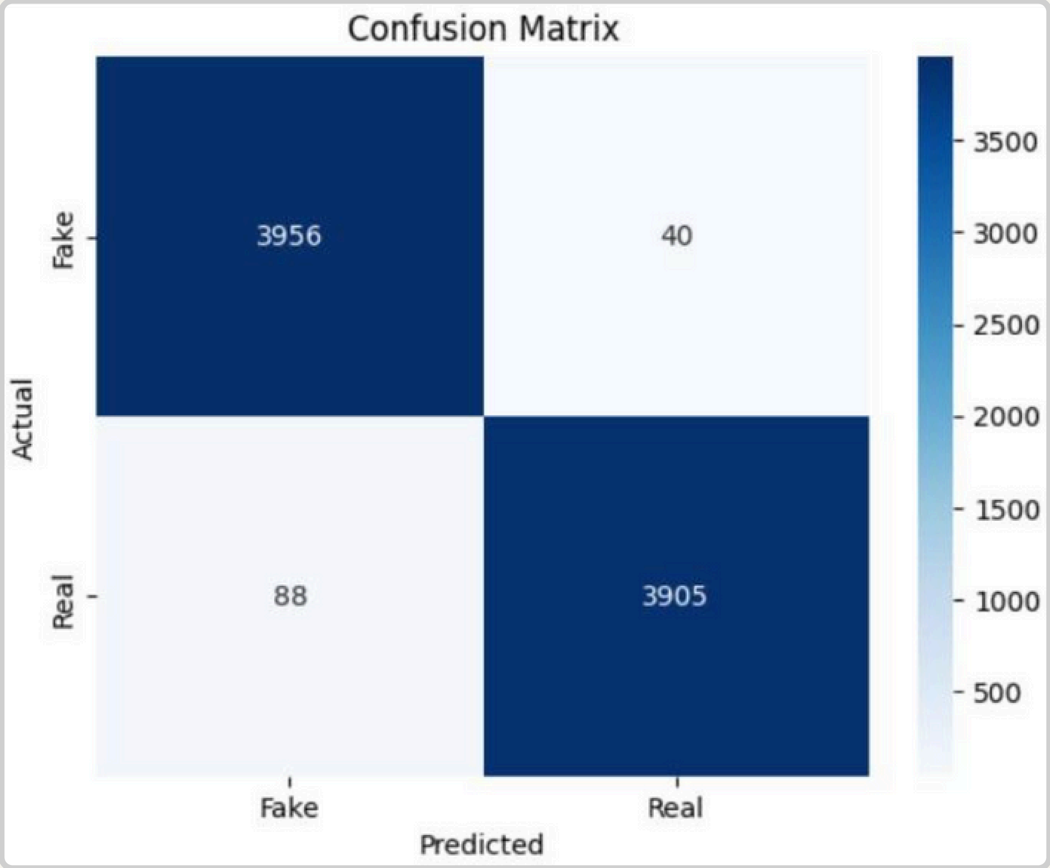
In the field of text classification, tasks such as spam detection, sentiment analysis, and topic categorization are common challenges. For these tasks, choosing the right classification model plays a significant role in achieving optimal results. Among the most widely utilized models are Logistic Regression, Multinomial Naive Bayes (MNB), and Support Vector Machines (SVM). These models are foundational in text-based classification due to their effectiveness, simplicity, and adaptability.

Classical ML

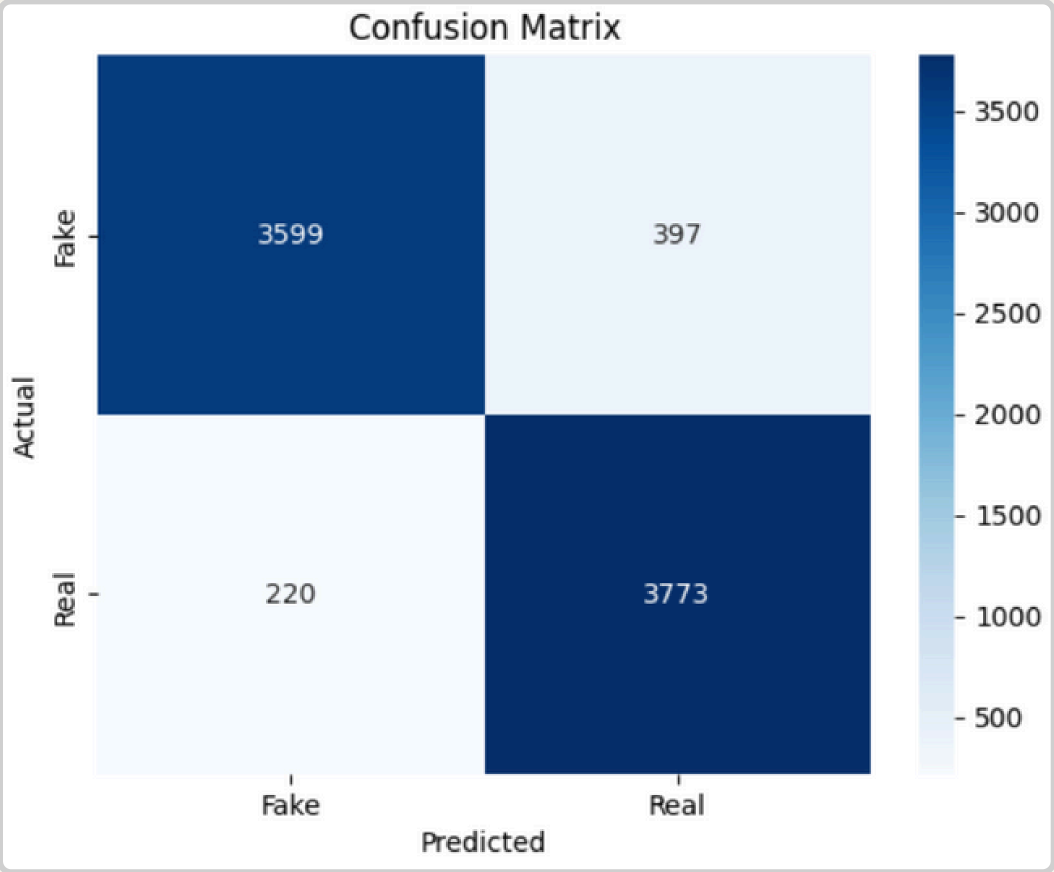
SVM	Multinomial Naive Bayes	Logistic Regression
supervised learning algorithm that finds the optimal hyperplane to separate data points into different classes by maximizing the margin between them.	probabilistic classifier based on Bayes' Theorem that assumes feature independence and calculates the likelihood of a class given word frequencies in text classification.	statistical model that estimates the probability of a class using a logistic (sigmoid) function
Accuracy: 98%	Accuracy: 92.28%	Accuracy: 98.87%

Classical ML

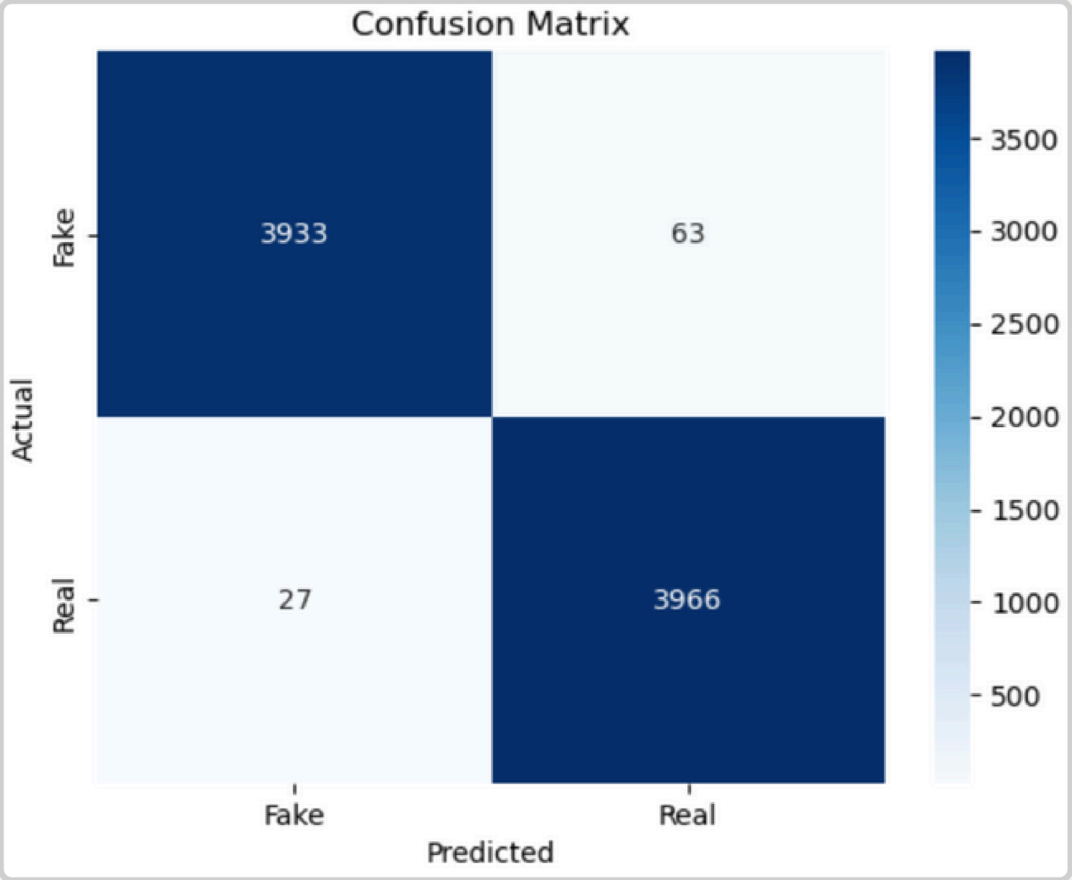
SVM



Multinomial Bayes



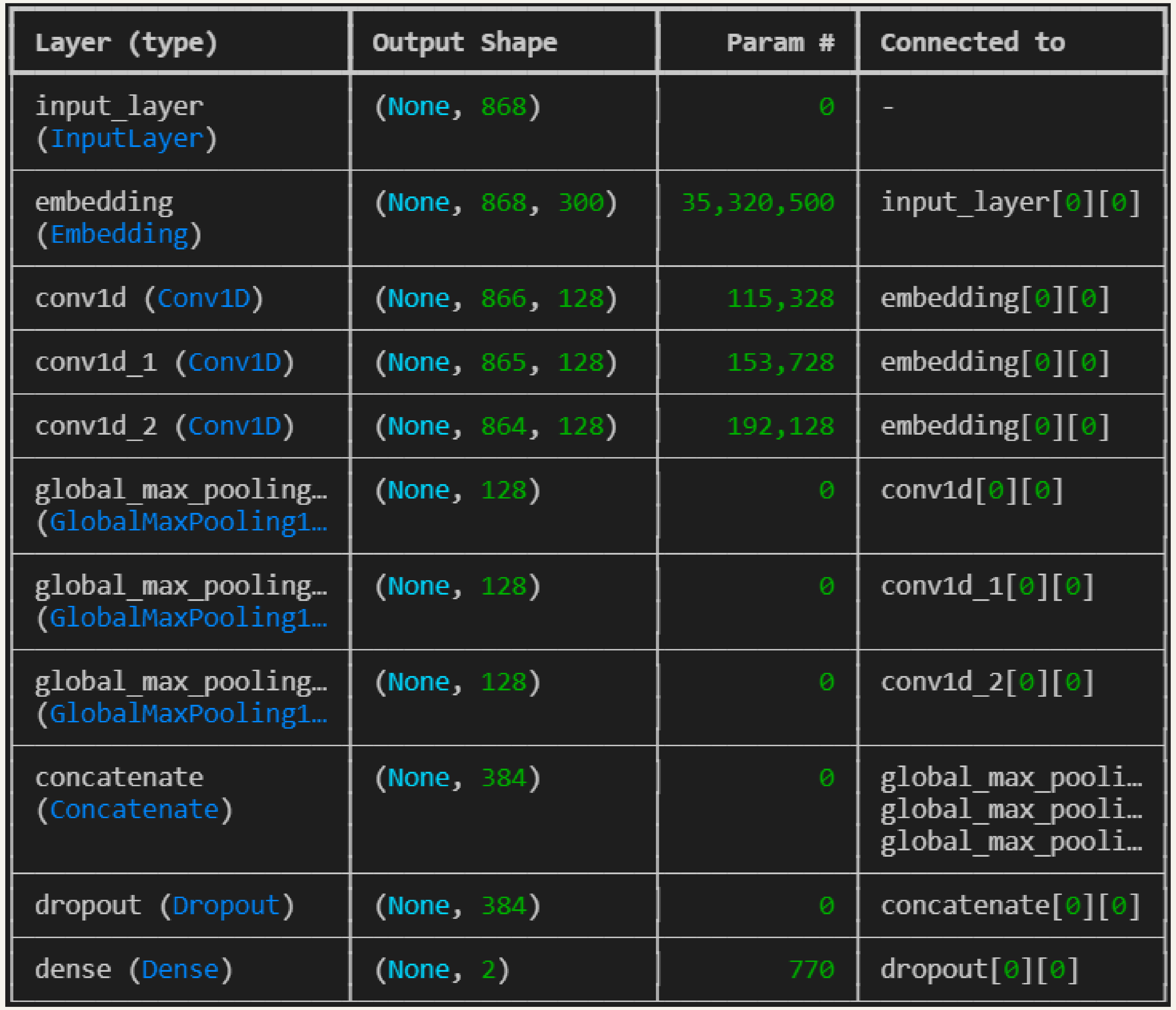
Logistic Regression



Overview: Kim-CNN

Kim-CNN is a method for understanding sentences by combining word embeddings with a convolutional neural network.

Kim-CNN Structure as shown:



Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 868)	0	-
embedding (Embedding)	(None, 868, 300)	35,320,500	input_layer[0][0]
conv1d (Conv1D)	(None, 866, 128)	115,328	embedding[0][0]
conv1d_1 (Conv1D)	(None, 865, 128)	153,728	embedding[0][0]
conv1d_2 (Conv1D)	(None, 864, 128)	192,128	embedding[0][0]
global_max_pooling… (GlobalMaxPooling1…	(None, 128)	0	conv1d[0][0]
global_max_pooling… (GlobalMaxPooling1…	(None, 128)	0	conv1d_1[0][0]
global_max_pooling… (GlobalMaxPooling1…	(None, 128)	0	conv1d_2[0][0]
concatenate (Concatenate)	(None, 384)	0	global_max_pooli… global_max_pooli… global_max_pooli…
dropout (Dropout)	(None, 384)	0	concatenate[0][0]
dense (Dense)	(None, 2)	770	dropout[0][0]

Training Process

Kim-CNN Hyperparameters:

- filter_sizes = [3, 4, 5]
- num_filters = 128
- dropout_rate = 0.5
- num_classes = 2

Training Hyperparameters:

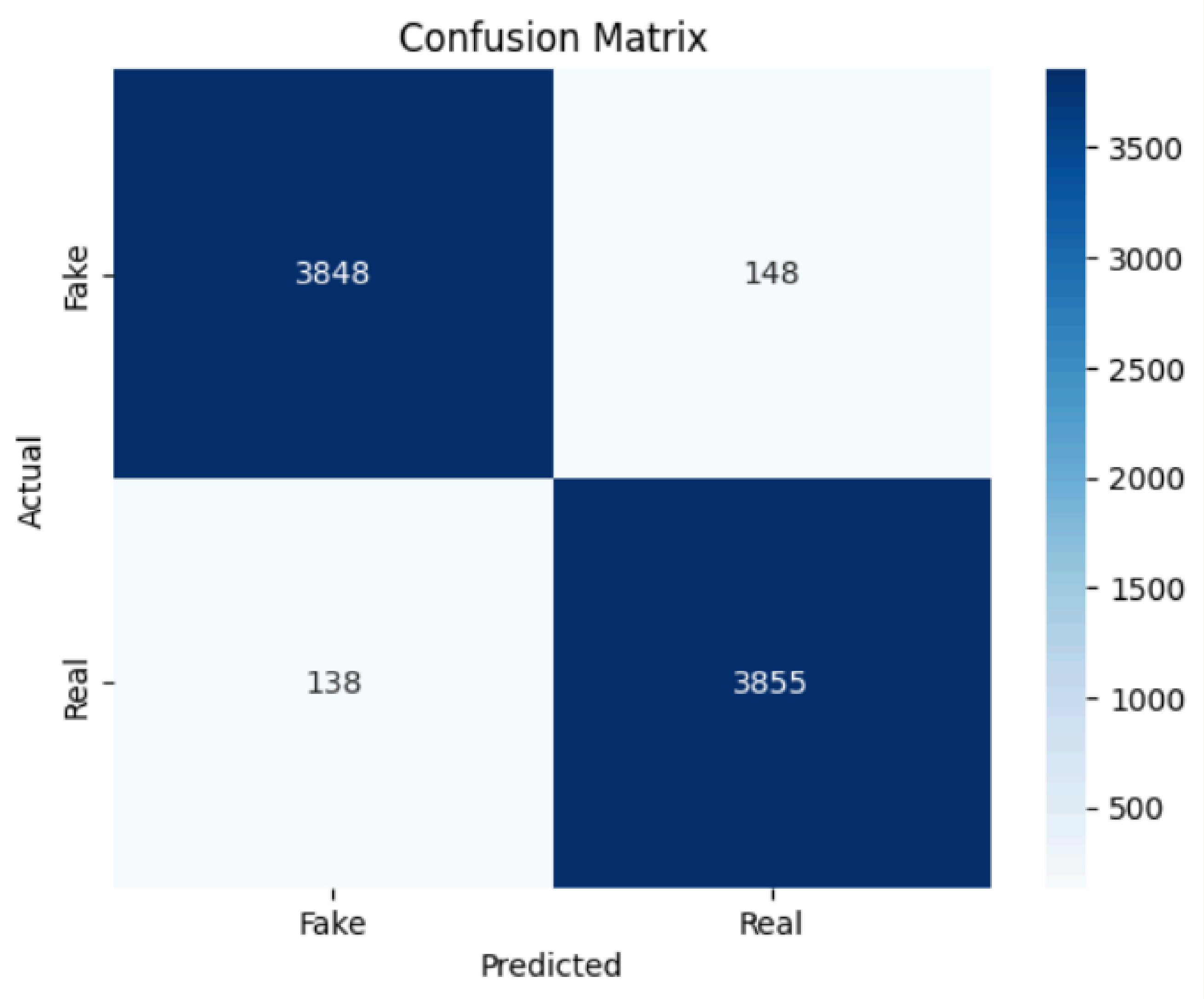
- batch_size = 128
- epochs = 10
- callbacks:

  - Stopping Early

  - Checkpoint
- Optimizer: Adam

Kim-CNN Result

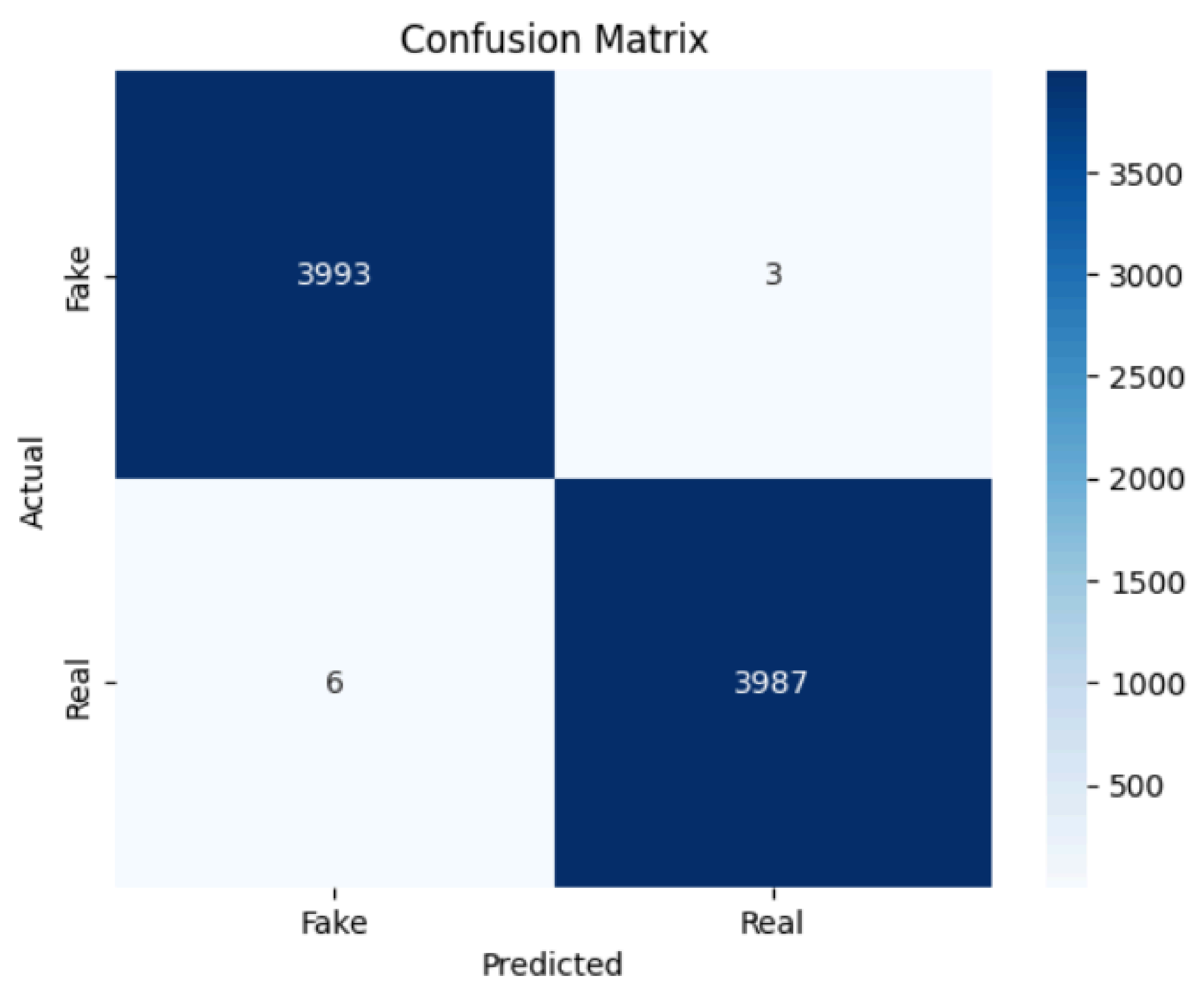
Training on The Title Column



Accuracy: 96.42%

Loss: 10.49%

Training on The Text Column



Accuracy: 99.88%

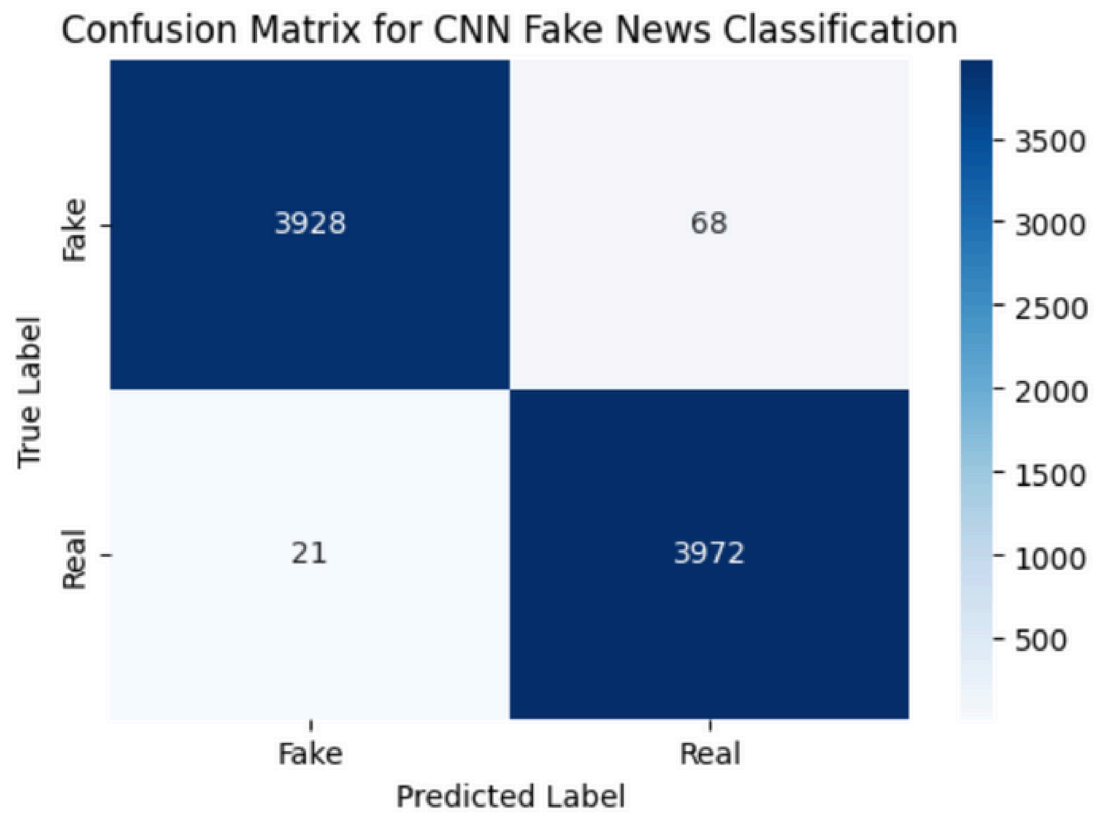
Loss: 0.0049%

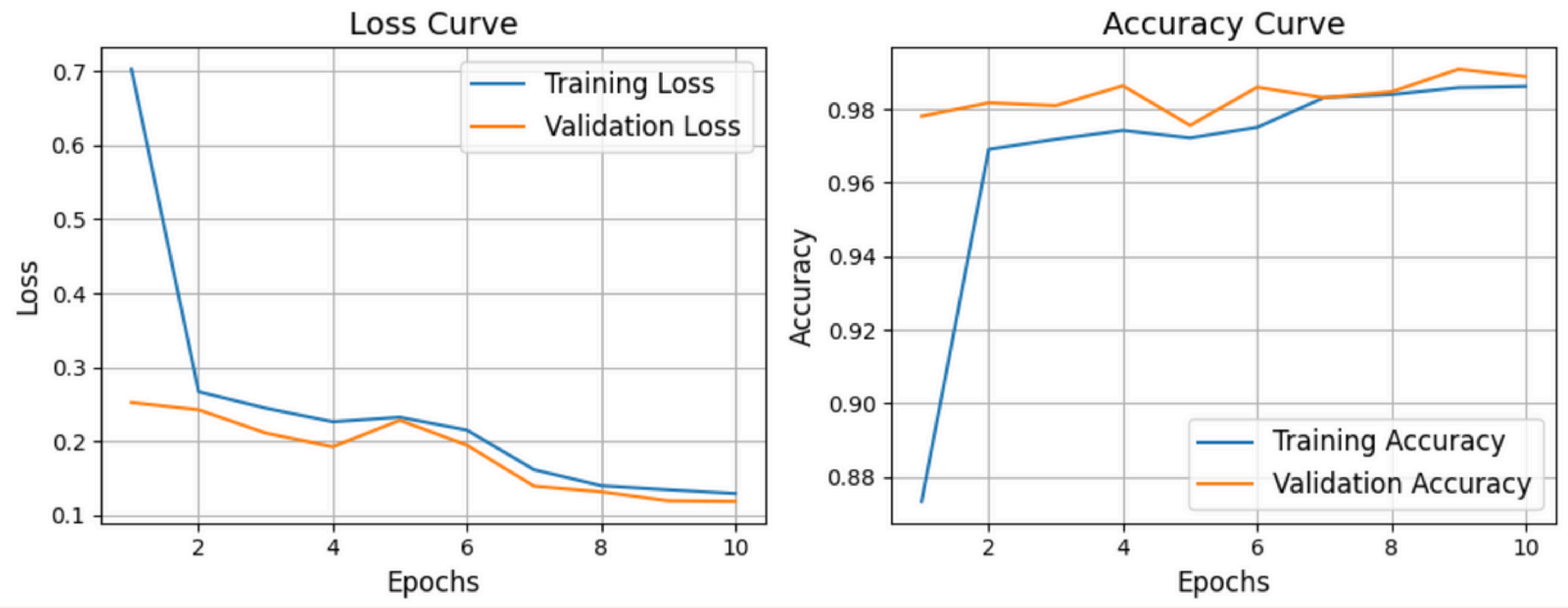
Other Model

- **CNN Architecture: Specifically designed for text classification tasks.**
- **Pre-trained GloVe Embeddings**: Enhance the model by providing semantic word representations, aiding in context and meaning comprehension.
- **Regularization Techniques:** Techniques like Dropout and L2 regularization help prevent overfitting and improve generalization.

Test Accuracy: 0.99

Test Loss: 0.12





Overall Result

Classical ML Models:

- SVM: 98% accuracy
- Multinomial Naive Bayes: 92.28% accuracy
- Logistic Regression: 98.87% accuracy

Kim-CNN Model:

- Title Column: 96.42% accuracy, 10.49% loss
- Text Column: 99.88% accuracy, 0.0049% loss

CNN with GloVe Embeddings:

- Test Accuracy: 99% , Loss: 0.12

Challanges:

- Selecting the best n-grams, embeddings (TF-IDF, Word2Vec) for classification is challenging.

DEMO Time!

LARA TEAM

IRON HACK

Thank you for listening

ANY QUISTIONS?

Project Repo:
https://github.com/LaylaZx/PROJECT-NLP-Challenge

Demo Link:
https://insyjix9kxiis3wihih6zg.streamlit.app/