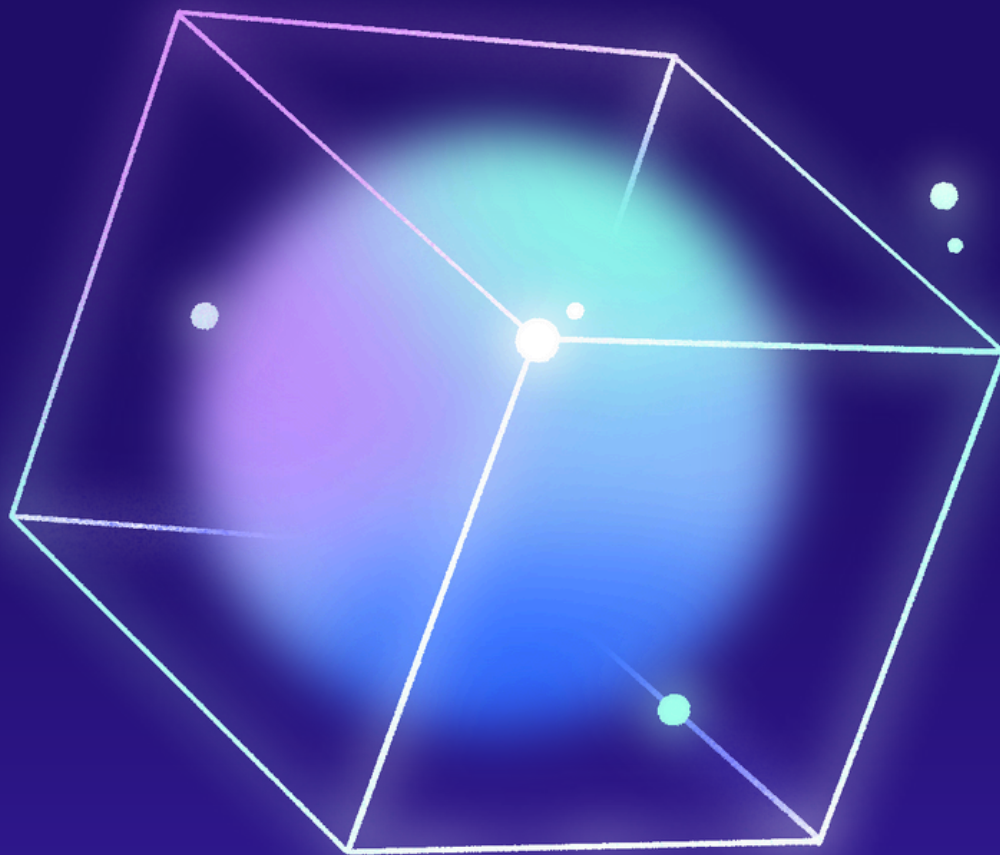


VERBASENSE

Amjad Althagafi, Taif Aladwani, Lama Alwuthaynani



PROJECT OVERVIEW

Goal:

Classify news headlines into two categories (0/1) using three distinct modeling approaches to compare performance and interpretability:

1. Traditional NLP Models:

- Logistic Regression with CountVectorizer/TF-IDF for baseline accuracy.
- Multinomial Naive Bayes to assess probabilistic term-class relationships.
- Support Vector Machine (SVM) to maximizes the margin between classes for potentially better generalization.

2. Deep Learning Model:

- Convolutional Neural Network (CNN) with word embeddings to capture local semantic patterns and spatial dependencies in headlines.

DATA PREPROCESSING

- Data Reduction: Sampled 20% of the data
- Feature Selection: Dropped irrelevant columns (date, subject, text), retaining only label and title.
- Lowercasing, removing punctuation & Stopwords
- Tokenization & lemmatization
- Feature extraction (TF-IDF, BOW)



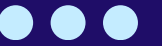
CLASSICAL NLP MODEL

Feature Extraction:

- BoW (Count Vectorizer)
- TF-IDF

Models Used:

- Logistic Regression
- Multinomial Naïve Bayes
- Linear SVM



PERFORMANCE OF CLASSICAL MODELS

Model	BoW Accuracy	TF-IDF Accuracy
Logistic Regression	91.68%	91.43%
Multinomial Naïve Bayes	91.61%	91.30%
Linear SVM	90.68%	91.30%

WORD2VEC-BASED CLASSIFIER

Word2Vec-based Classifier for Fake News Detection

Goal:

Train a classifier using Word2Vec embeddings to classify fake and real news based on the article title instead of the full text.

1 Model Training on Titles Only

- Instead of using full article text, the model was trained on titles only.
- This reduces the amount of text while keeping key information.

2 Results Comparison

- Classical Model (Logistic Regression): 91% accuracy.
- CNN Model (Kim's CNN): 96% accuracy, with no overfitting!

3 Why is Overfitting No Longer an Issue?

- Titles are shorter, reducing the risk of capturing unnecessary patterns.
- Using Dropout and a larger Batch Size improved model stability.

DEPLOYING WITH STREAMLIT

News Headline Classification App

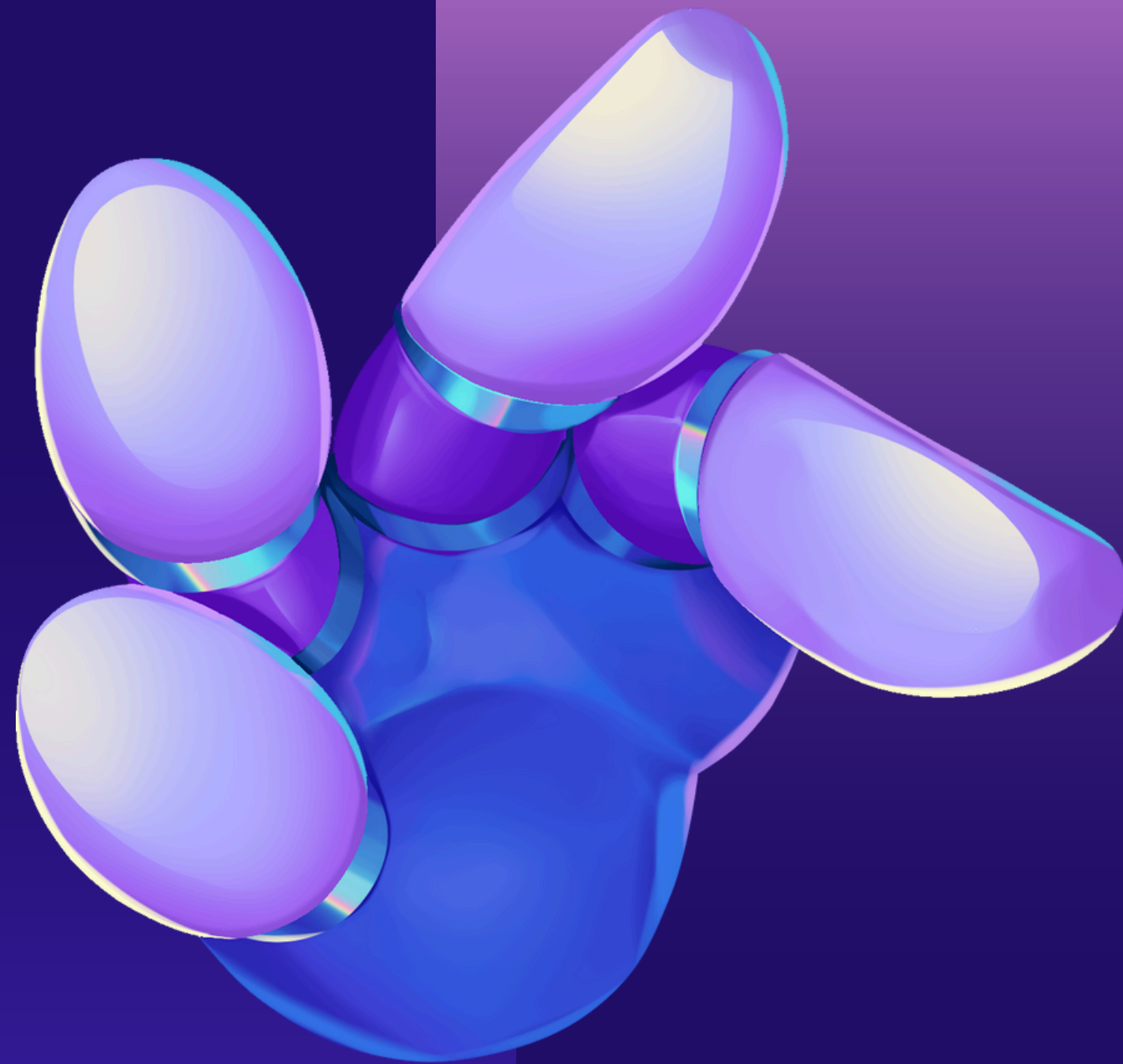
Enter a news headline to classify it using the trained CNN model.

Enter a headline:

U.S. House panel approves broad proposal on self-driving cars

Predict

Prediction: Positive (Confidence: 0.9959)



This project successfully demonstrated that raw headline text, without heavy preprocessing, can provide sufficient signal for accurate classification into predefined categories. By focusing on word-based features and applying vectorization techniques such as CountVectorizer and TF-IDF, we were able to highlight the inherent discriminative power of raw headlines. Logistic Regression emerged as the most effective model, proving its ability to adapt to and learn from unprocessed text data.

CONCLUSION

THANK YOU

