Sarcasm Detection In Movie Reviews

Sarcasm, a form of verbal irony where the intended meaning is opposite to the literal words used, can obscure the true sentiment in movie reviews. Detecting sarcasm is crucial for accurately interpreting the emotional tone and overall opinion expressed in the reviews.

by Nikhitha Kurumaddali

GROUP 6

BATCH 02
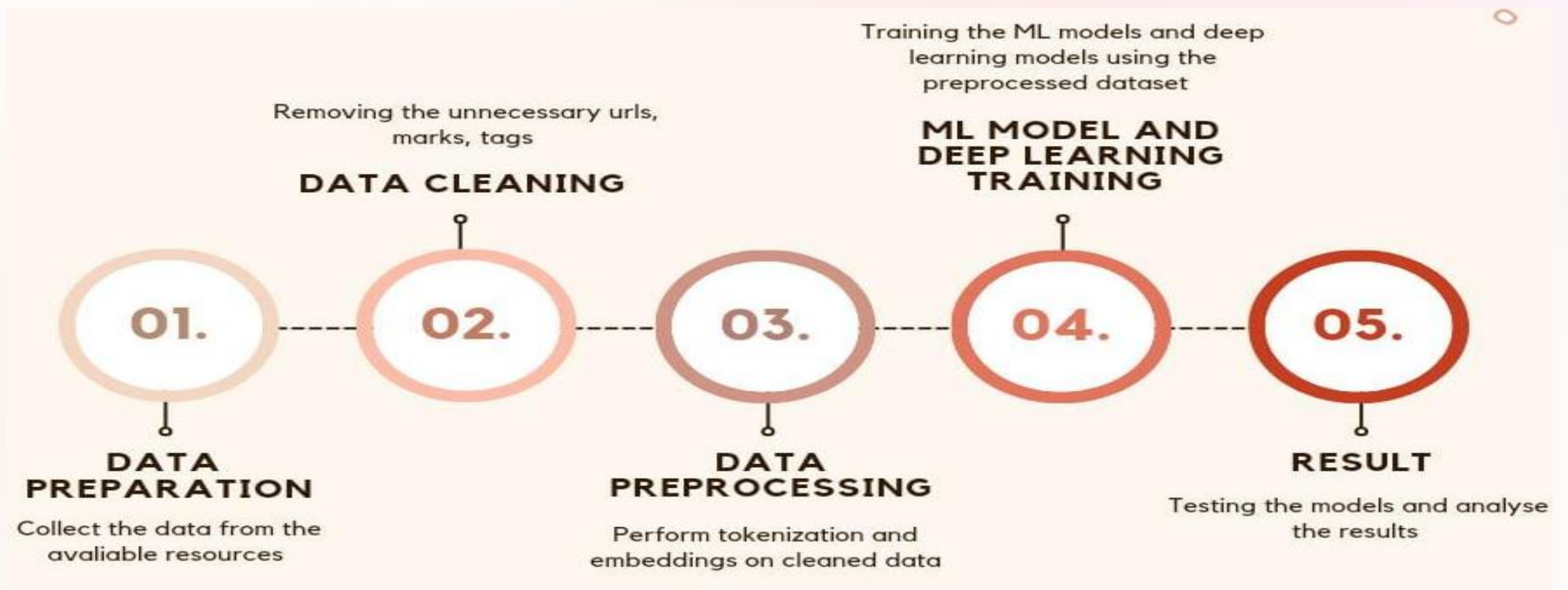
Infosys Springboard internship

Business Problem in Movie Reviews

- **Movie success hinges on audience feedback**, making review analysis crucial for accurate ratings.

- Websites like BookMyShow, IMDb, and Rotten Tomatoes are pivotal in collecting and presenting audience reviews.

- However, **some reviewers express their thoughts sarcastically**, which can mislead traditional sentiment analysis models. This poses a challenge because **sarcasm can mask the true sentiment of a review**, leading to inaccurate analysis and potentially flawed ratings.

- **Positive Review:** "*Interesting, good filmmaking, surprisingly authentic.*" (Clear and positive)

- **Negative Review:** "*Dialogue is atrocious, acting phoned-in. Big disappointment.*" (Stronger wording emphasizes negativity)

- **Sarcastic Review (Difficult to Detect):** "*This movie was so bad, I'd rather gnaw on a golf ball.*" (Exaggerated statement hints at sarcasm)

Proposed Solution

- . To address the challenge of sarcasm detection, we need to develop and **implement** advanced **Deep Learning models** specifically designed to recognize and **interpret sarcastic reviews.**

- To achieve this, we will follow the **workflow outlined below to build the model**:

Training the ML models and deep learning models using the preprocessed dataset

ML MODEL AND DEEP LEARNING TRAINING

Removing the unnecessary urls, marks, tags

DATA CLEANING

01.   02.   03.   04.   05.

DATA PREPARATION

Collect the data from the avaliable resources

DATA PREPROCESSING

Perform tokenization and embeddings on cleaned data

RESULT

Testing the models and analyse the results

1) Data Preparation

1 **Data Collection**

We have **gathered movie reviews from popular online platforms**, such as IMDb, Rotten Tomatoes ensuring a diverse and representative dataset and labelling them as sarcastic and not sarcastic using online prediction models.

2 **Data Exploration**

We will perform an in-depth analysis of the dataset, **examining the distribution of review lengths, sentiment,** to gain a comprehensive understanding of the data.

3 **Handle the Reviews**

The dataset contained few **duplicate reviews rows** which are been **removed** . Also **dropped unnecessary columns**

4 **Handle the Labels**

The labels column contains few null values ,which are handled by **removing the rows with null value label**. The labels are **encoded** as **'0' for Not Sarcastic** and **"1" for Sarcastic**

5 **Split the dataset**

The dataset is **split into train data , test data and validation data** which are used for training and testing the ML and Deep Learning models

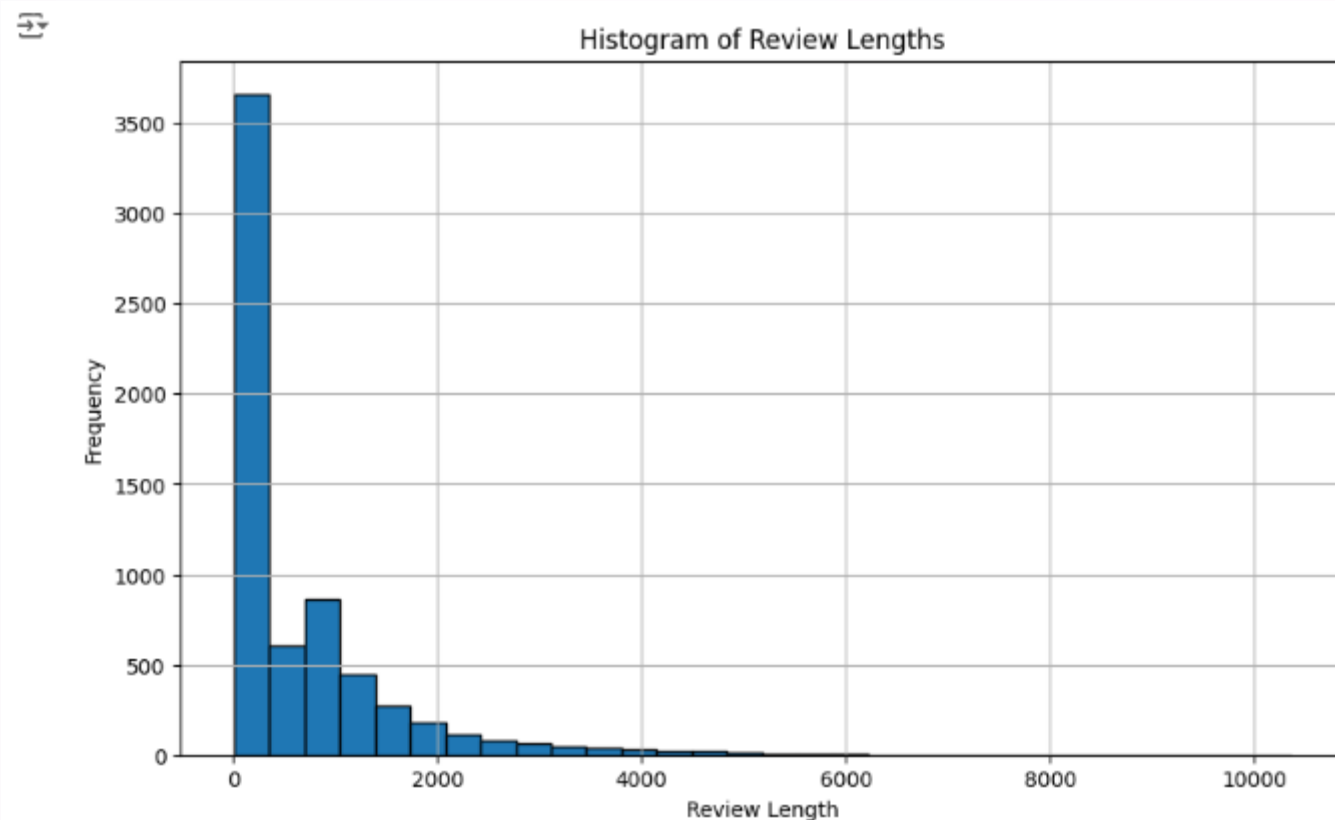| reviews	Sarcasm
One of the other reviewers has mentioned that ...	not sarcastic
A wonderful little production. The...	not sarcastic
This movie was a groundbreaking experience!<br...	sarcastic
I thought this was a wonderful way to spend ti...	not sarcastic
Basically there's a family where a little boy ...	sarcastic

```
Train datset size : 4157
Test datset size : 1300
validation datset size : 1040
```


Distribution of Data across Train, Test, and Validation Sets
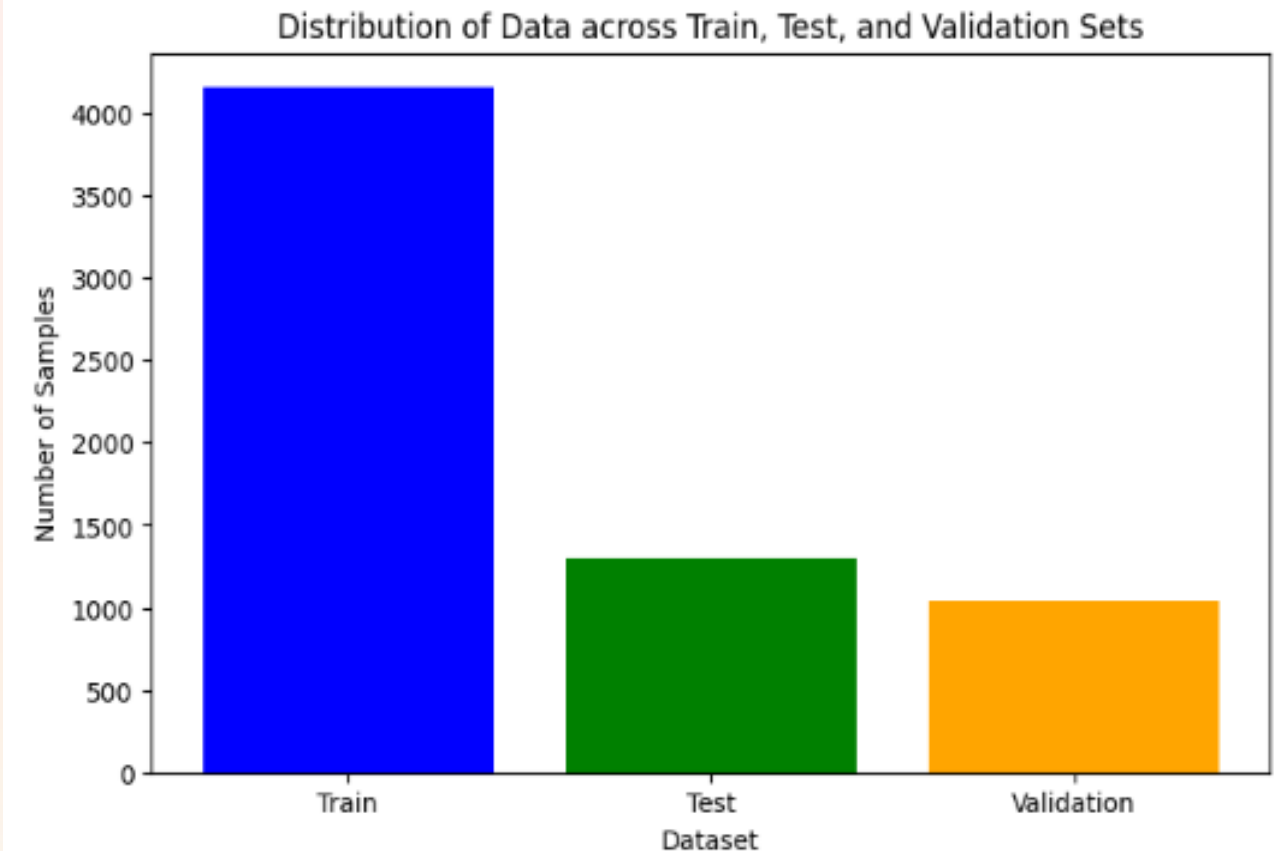
```
Mapping of sarcasm labels to numerical values:
not sarcastic: 0
sarcastic: 1
```

Insights on the Dataset

- We have chosen the **IMDB movie reviews** dataset for this project **from Kaggle**, it contains a large volume of reviews, offering a **diverse range of sentiments and writing styles**.

- The dataset contains **around 7000 reviews labelled as Sarcastic and Not Sarcastic**

- No.of reviews that are **Sarcastic : 3918**   No.of reviews that are **Not Sarcastic : 2979**

- The **mean length of review** in the dataset is **655** words

Review Length
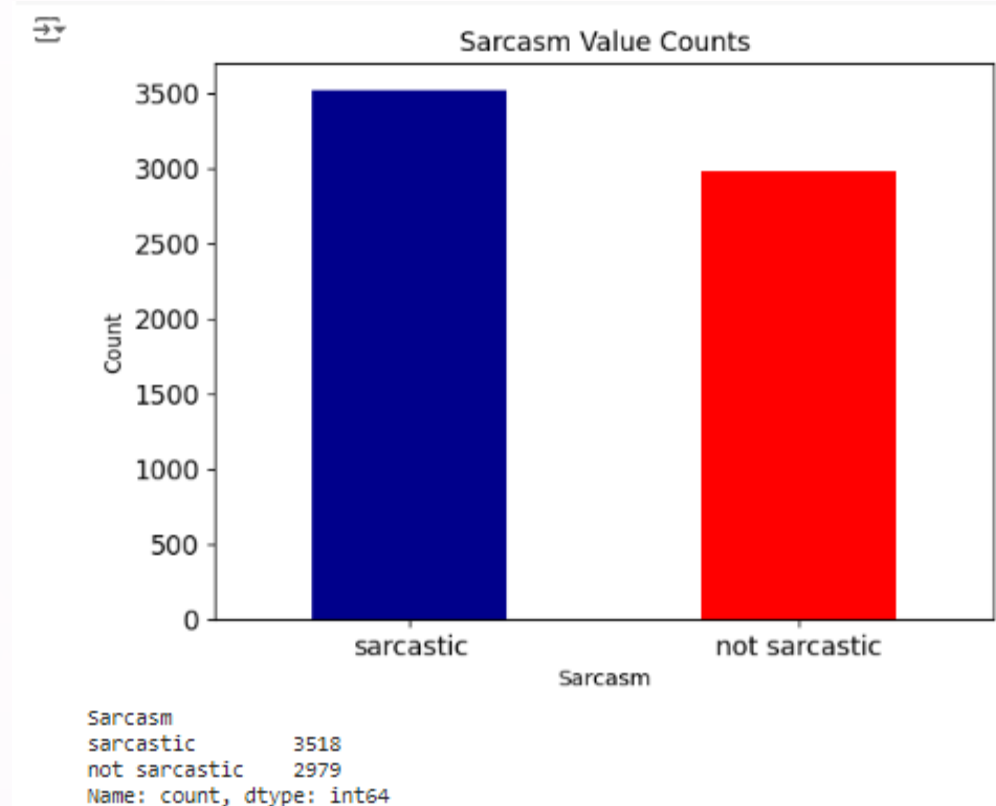
Sarcasm Distribution

2) Data Cleaning

Text Normalization

1 We will standardize the review text by removing HTML tags, and eliminate unnecessary punctuation marks that do not contribute to sentiment or sarcasm detection. Removing any blank lines and excess white spaces

Stopword Removal

2 Common stopwords, such as "the," "a," and "is," will be removed to focus the model's attention on more informative words.

Lemmatization

3 Words will be reduced to their base forms (lemmas) to improve the model's understanding of semantic relationships and reduce vocabulary size.

cleaned_reviews	lemmatized_reviews
One of the other reviewers has mentioned that ...	One reviewer mention watch 1 Oz episode youll ...
A wonderful little production. The filming tec...	wonderful little production. filming technique...
This movie was a groundbreaking experience! Iv...	movie groundbreaking experience! Ive never see...
I thought this was a wonderful way to spend ti...	think wonderful way spend time hot summer week...
Basically theres a family where a little boy J...	Basically there family little boy Jake think t...

3) Data Preprocessing

A)   Tokenization

The review text will be **broken down into individual words or tokens**, which will serve as the basic units for the deep learning model.For our model we have **used Keras Tokenizer**

B)   Vocabulary Building

A comprehensive **vocabulary will be constructed**, including the **most frequent words** in the dataset, to represent the diverse language used in movie reviews.

C)   Embedding

The **tokens will be transformed into numerical representations** (embeddings) that capture the semantic relationships between words, enabling the model to understand the contextual meaning of the reviews.

D)   Padding

To ensure **all input sequences have the same length**, the reviews will be padded with a special token, allowing the model to process batches of data efficiently. For our model we have **used Keras Pad Sequences**

4) **Model Training**

- Initially, we applied traditional ML models including **logistic regression, Naive Bayes, decision tree, and Random Forest.** After hyperparameter tuning, **Random Forest emerged as the top** performer due to its ability to handle complex relationships in the data.

- Transitioning to deep learning, we explored architectures such as **Neural Networks (NN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and BERT models.**
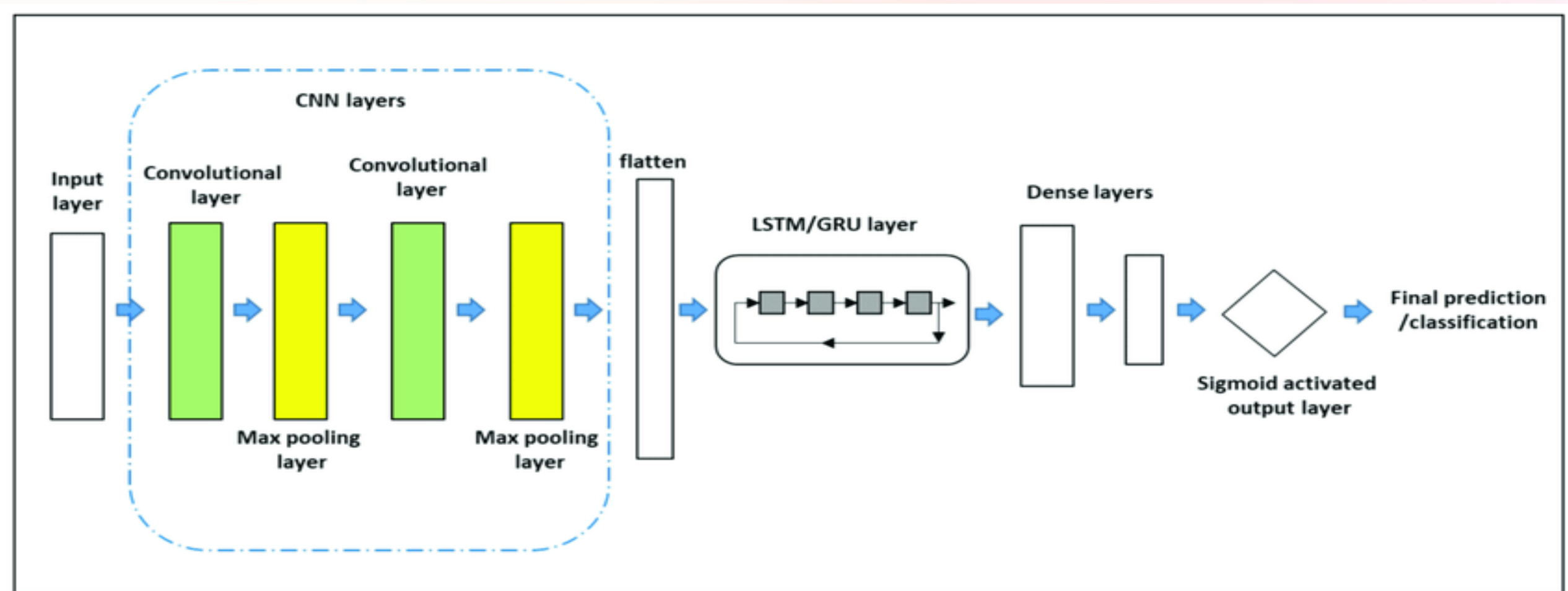
Why GRU Model for our dataset ?

- GRUs are a **type of recurrent neural network (RNN)** that excel at modeling sequential data.
- They maintain an internal state that **captures dependencies** across time steps, allowing them to **learn context and patterns** in text.
- For sarcasm detection, this ability to capture sequential information is crucial because **sarcasm often relies on subtle cues spread across multiple words or phrases.**
- Combining **Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs)** in a hybrid model can leverage their complementary strengths.
- **CNN**s excel at **capturing local patterns** (e.g., n-grams), while **GRU**s handle **sequential dependencies**.
- By combining these architectures, we can **enhance** sarcasm detection **accuracy**.

```
Model: "sequential"
_____
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 942, 128)          3781888

 conv1d (Conv1D)             (None, 938, 128)          82048

 max_pooling1d (MaxPooling1  (None, 469, 128)          0
 D)

 bidirectional (Bidirection  (None, 469, 1024)         1972224
 al)

 bidirectional_1 (Bidirecti  (None, 256)               886272
 onal)

 dense (Dense)               (None, 128)               32896

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 32)                2080

 dense_3 (Dense)             (None, 1)                 33
```

5) Results and Constraints

- The hybrid **CNN-GRU model** demonstrates **promising performance** in sarcasm detection.
- **High precision (0.90) for class 1** (sarcasm) indicates **accurate predictions** when identifying sarcastic expressions.
- **Balanced recall (0.75)** ensures that the model **captures a significant portion** of actual sarcasm instances.
- The GRU-based model demonstrates **strong performance**, achieving an overall **accuracy of 82%** on the dataset, with **balanced precision and recall** across both classes.

	precision	recall	f1-score	support
0	0.76	0.91	0.83	604
1	0.90	0.75	0.82	696
accuracy			0.82	1300
macro avg	0.83	0.83	0.82	1300
weighted avg	0.84	0.82	0.82	1300

CONSTRAINTS

- The project **does not include** detecting sarcasm in **non-English reviews**.
- It **does not handle non-textual sarcasm** such as in images, videos, or memes.
- Real-time detection of sarcasm during review submission is also not covered.

Real-time Application and Future Scope

Real-time Deployment

This model can be **applied to platforms like Rotten Tomatoes**, which rely on user reviews to rate movies. By integrating sentiment analysis and our sarcasm detection model, we can enhance the accuracy of review ratings. This approach **mitigates issues such as human bias, reviewer subjectivity, and language nuances, providing more reliable insights into audience sentiments**.

Future Enhancements

Future enhancements could focus on distinguishing between **sarcastic positive and sarcastic negative reviews.** This advancement would further refine the movie review analysis system, **offering deeper insights into audience reactions** and improving decision-making for filmmakers and critics alike.

Conclusion

In conclusion, we have **developed a sarcasm detection model for movie reviews** using the **IMDB dataset from Kaggle**. The dataset underwent thorough **cleaning and preprocessing.** Initially, the preprocessed data was applied to a **Random Forest ML model,** achieving (mention precision and recall values). To enhance detection accuracy, we transitioned to a deep learning model, specifically the **GRU model,** which resulted in **good performance.** This approach effectively **addresses ambiguity in movie review analysis caused by sarcastic expressions.**

Thank You