Comment Filtering System

—— Group 9

Content

PART 01

Introduction

Introduction

Problem Statement:

- Comments may contain illegal text (pornographic, abusive)
- Large volume of comments makes it difficult to identify high-quality and meaningful insights.

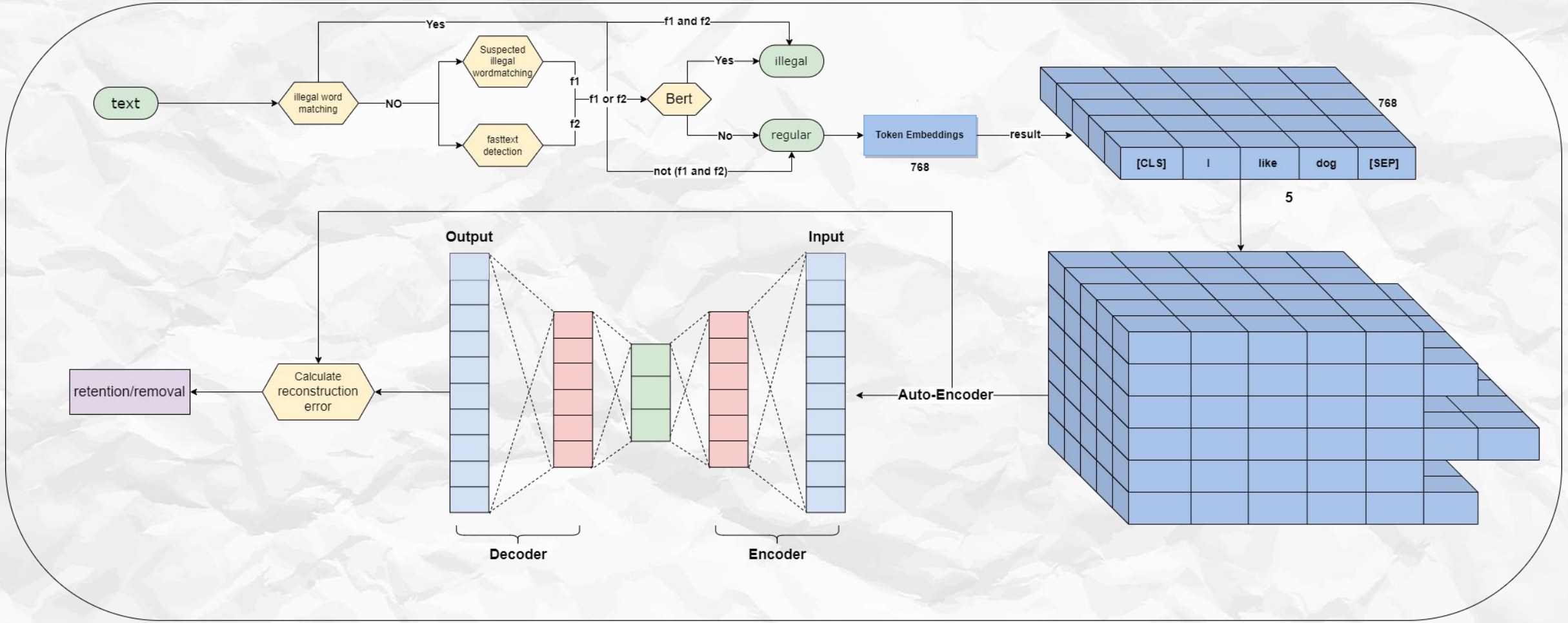
Project Goals:

- Filter then retain high quality comments.

PART 02

Project Overview

Project Workflow



PART 03

Illegal Text Detection

Data Collection

Collected Data:

- The dataset consists of approximately 19,470 text samples, categorized as binary classification data (harmful vs. harmless text).
- The harmful category includes content such as hate speech, sexual offenses, offensive humor, and sarcasm.

Harmful Texts:

- **Hate Speech:**
  Targeting a specific group of people, such as racial slurs.
- **Sexual Content:**
  Violent, vulgar language aimed at those over 18.
- **Satire and Hell Jokes:**
  prevalent on social media, such as offensive memes.

Data Source: Github

- COLDataset
- ToxiCloakCN

Data Augmentation

❖ Data Augmentation Methods

- **Insert irrelevant characters.**

  For example: “变#,,态”

- **Break the word.**

  For example: “亻 尔”

- **Homophone substitution.**

  For example: “卧槽”

- **Pinyin substitution**:

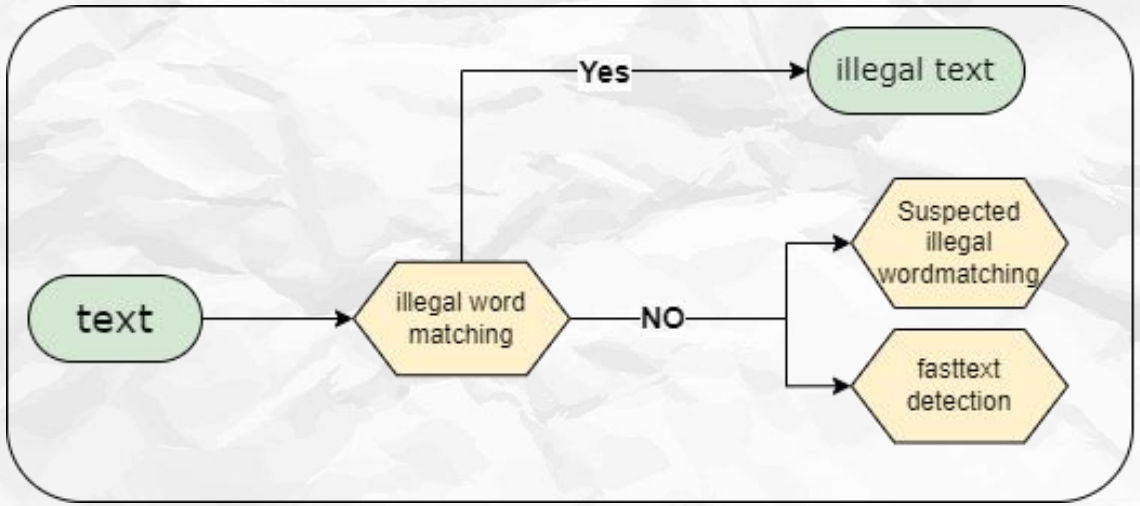
  “wocao”

Illegal word matching

➢Method :

Aho-Corasick AutoMaton is a string matching algorithm used to solve multi-pattern matching.

- Brute Force Algorithm $O(|S| \cdot |P|)$
- Aho-Corasick AutoMaton Algorithm $O(|P| + |S|)$



➢Example:

That nigga is rude!

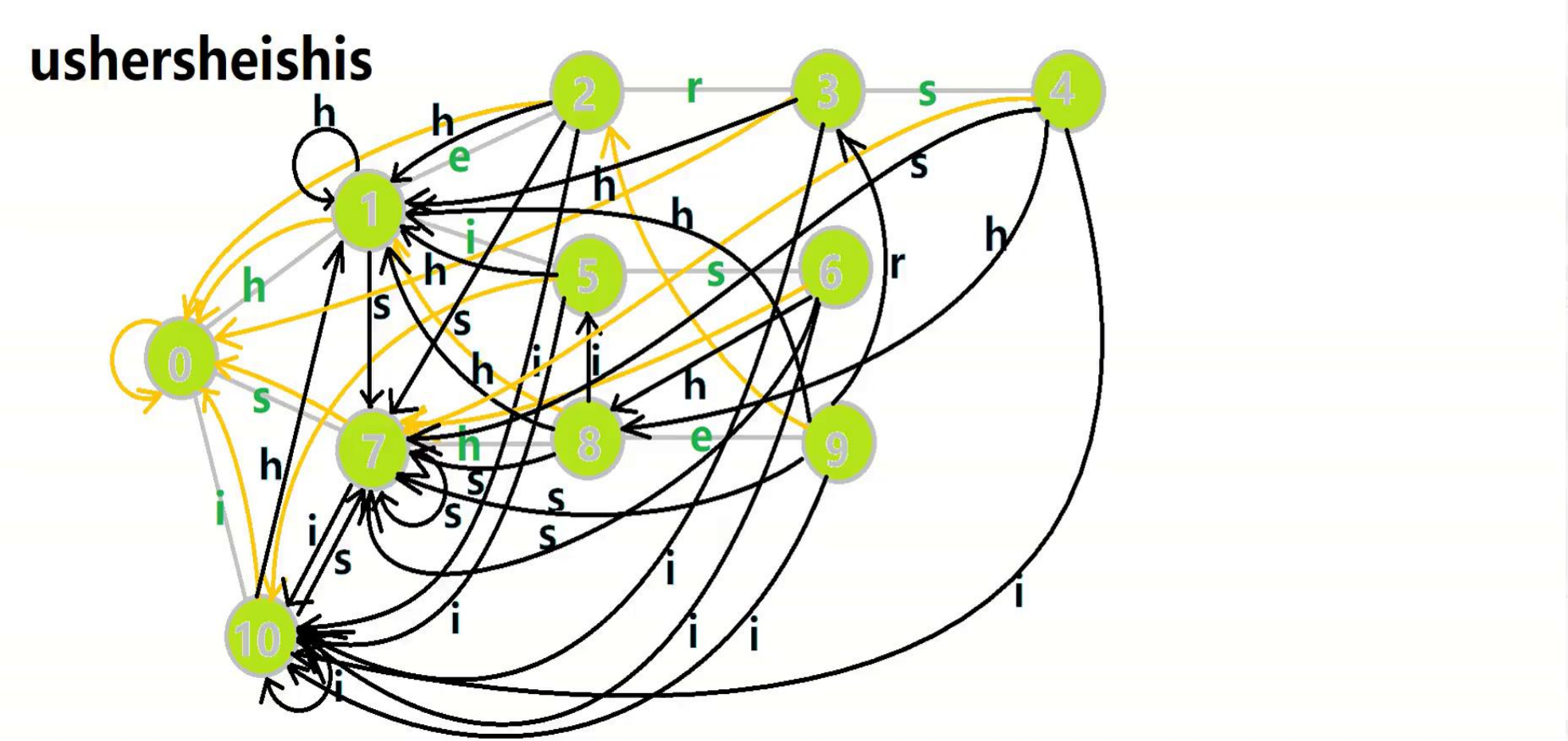
Match by Aho-Corasick AutoMaton

That **nigga** is rude!

Illegal text!

Illegal word matching

Building an Aho-Corasick AutoMaton which can match words : i, he, his, she, hers , and handling a sentence “usherssheishis”



Text Classification Machine Learning Algorithm

Machine Learning Methods

- **Multinomial Navie Bayes**
- **Random Forest**
- **XGBoost**
- **LightGBM**
- **SVM**
- **Logistic Regression**

Model	Accuracy	Precision	Recall	F1-Score
MultinomialNB	0.5461	0.5427	0.5461	0.5441
Random Forest	0.5899	0.5905	0.5899	0.4942
XGBoost	0.5847	0.5672	0.5847	0.5165
LightGBM	0.5784	0.5548	0.5784	0.5086
SVM	0.5680	0.5491	0.5680	0.5423
Logistic Regression	0.5600	0.5413	0.5600	0.5375

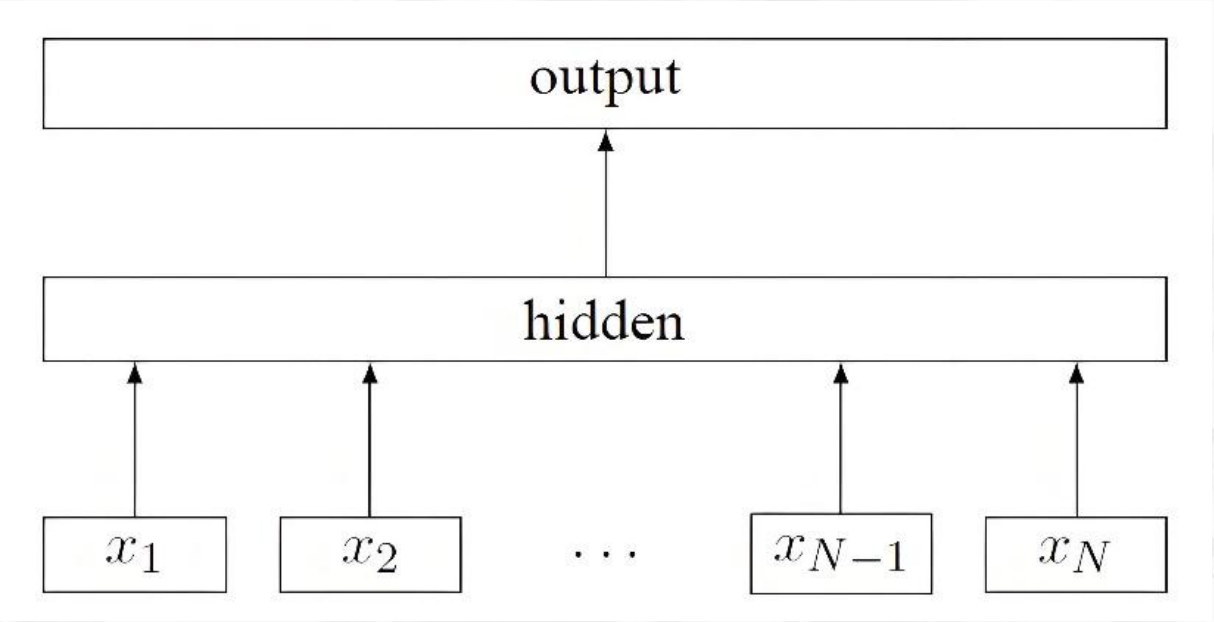
Performance were not satisfactory!

Text Classification by using deep learning method

- FastText
- TextCNN
- LSTM
- BERT

FastText

In traditional word embeddings, each word is represented as a single vector.
However, FastText improves on this by representing each word as a bag of **character n-grams**. This approach allows FastText to handle **out-of-vocabulary (OOV)** words better by using subword information to infer the meaning of new words based on their character composition.



For example: The word "apple" would be broken down into **n-grams** like:

- n=3: "app", "ppl", "ple"
- n=4: "appl", "pple"

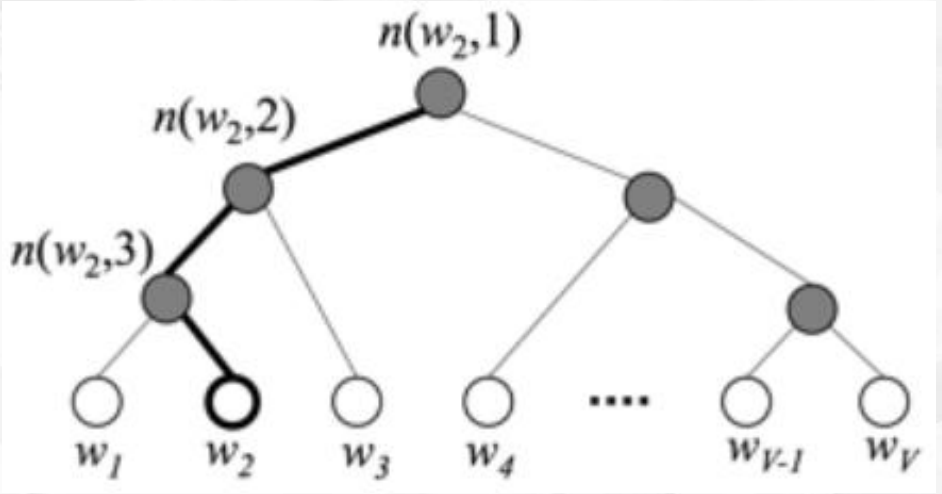
Advantage:

1. Handling Out-of-Vocabulary (OOV) Words
2. Efficient and Fast
3. Interpretability

FastText

Hierarchical Softmax:

In hierarchical softmax, the vocabulary is arranged in a **binary tree**, and words are predicted by traversing the tree. This significantly **reduces the computational complexity** compared to standard softmax.



N-gram:

n-gram is an algorithm based on language model, and the basic idea is to slide the text content into a window of size N in the order of subsections, and finally form a sequence of byte fragments with a window of N.

FastText Classification Report:

	precision	recall	f1-score	support
0	0.99	0.99	0.99	4112
1	0.99	0.98	0.98	3025
accuracy			0.99	7137
macro avg	0.99	0.99	0.99	7137
weighted avg	0.99	0.99	0.99	7137

BERT(Bidirectional Encoder Representations from Transformers)

Use a **Transformer** architecture, designed for natural language understanding tasks. It outputs vectors. Unlike traditional unidirectional language models (LSTM), BERT is **bidirectional**, meaning it considers **both the left and right context** of a word during training. This bidirectional nature enables BERT to better understand the **deep semantics** of words and sentences.

In our project, after we input the text, we will get vectors. Then those vectors will be passed to a **fully connected layer** to do **binary classification problem**.

Advantage: 1.Powerful Contextual Modeling
2.Task Flexibility
3.Efficient pretraining and Fine-tuning

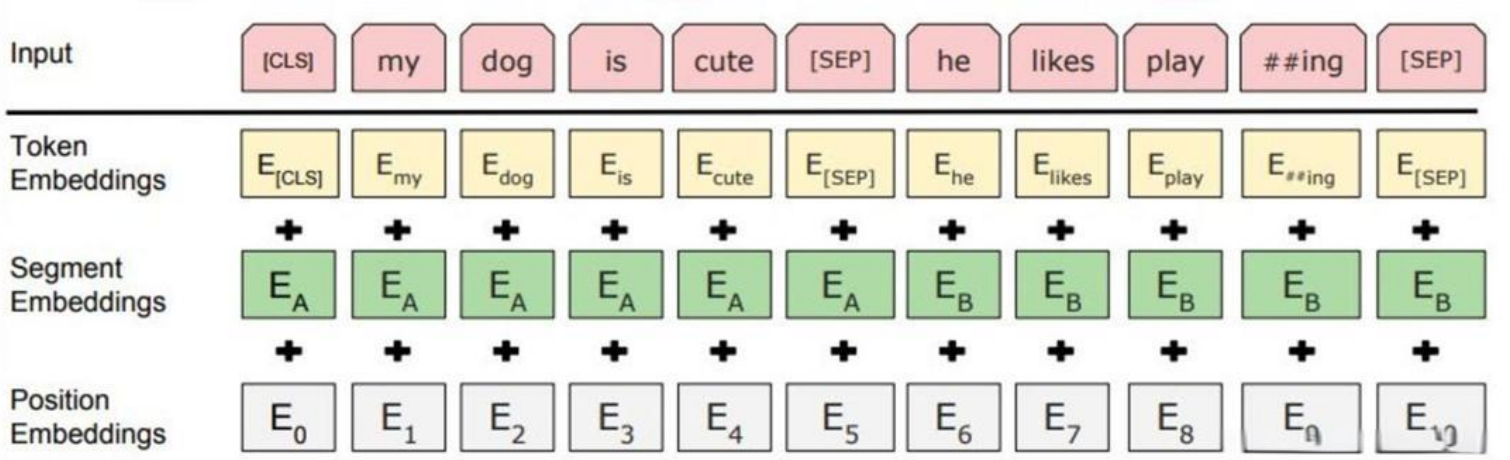


图2：BERT输入表示。输入嵌入是token embeddings, segmentation embeddings 和position embeddings 之和。

```
Classification Report:
              precision    recall  f1-score   support

     Class 0       1.00      0.98      0.99      4112
     Class 1       0.97      1.00      0.99      3025

    accuracy                           0.99      7137
   macro avg       0.99      0.99      0.99      7137
weighted avg       0.99      0.99      0.99      7137
```

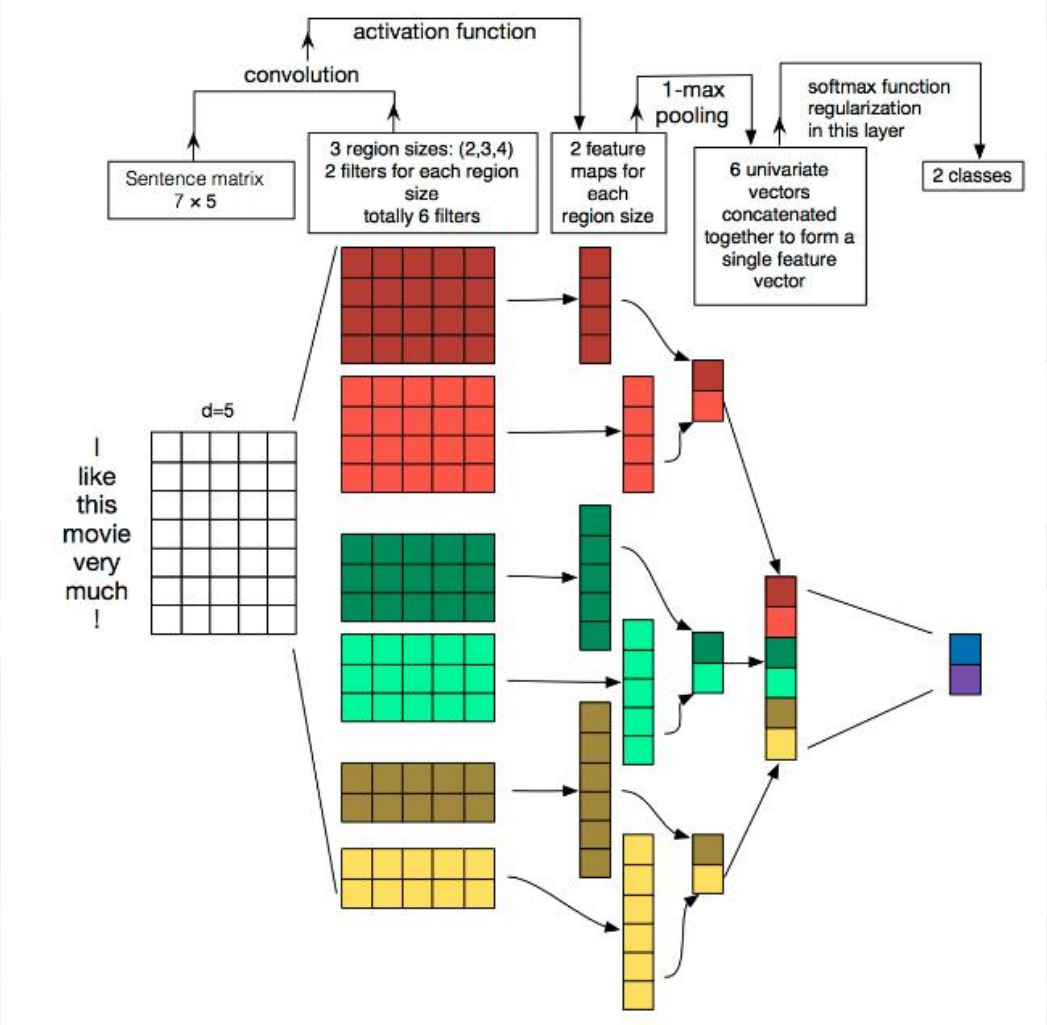
TextCNN

TextCNN
uses convolution operations to extract local features from text, followed by pooling layers to reduce dimensionality, and then a fully connected layer for classification. It has four main steps:

1. Embedding Layer
2. Convolution Layer
3. Pooling Layer
4. Fully Connected Layer

Advantage:

1. Efficiency and Speed
2. Simplicity and Ease of Use
3. Handling Long Texts



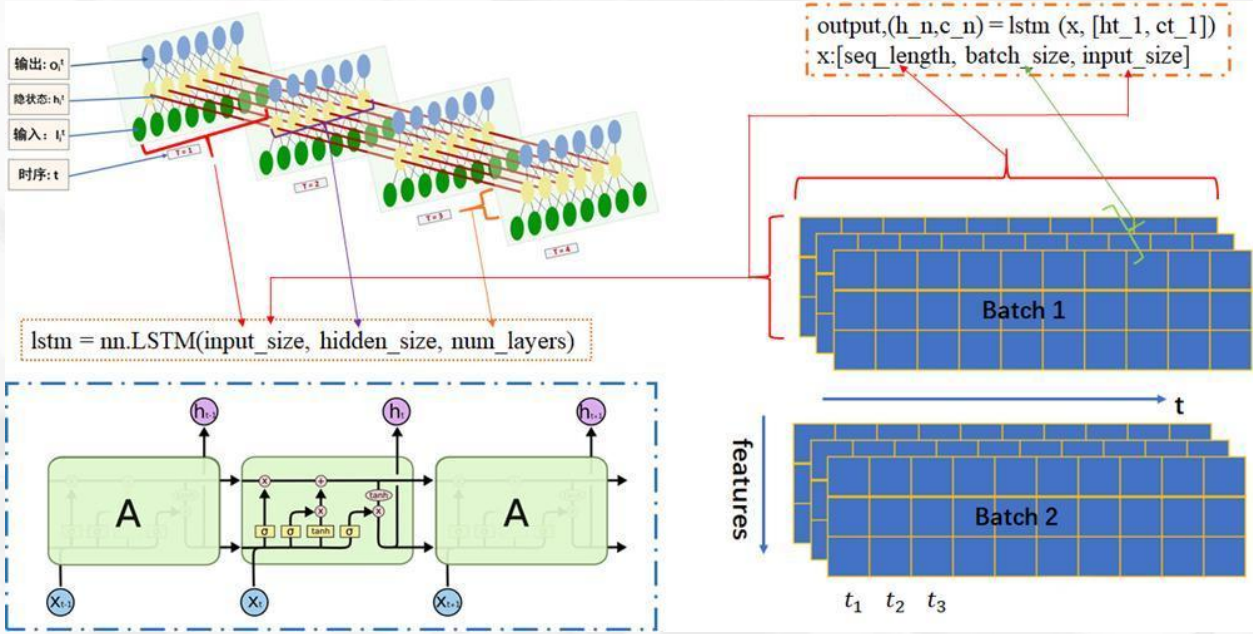
Classification Report:	precision	recall	f1-score	support
Class 0	0.99	0.99	0.99	4112
Class 1	0.99	0.98	0.99	3025
accuracy			0.99	7137
macro avg	0.99	0.99	0.99	7137
weighted avg	0.99	0.99	0.99	7137

LSTM

LSTM is designed to address the vanishing gradient and exploding gradient problems in traditional RNNs when handling long sequences.
LSTM introduces gating mechanisms that allow it to retain and forget information selectively, enabling it to capture long-term dependencies in sequence data.

It has the following four key components:

1. Forget Gate:
2. Input Gate
3. Cell State
4. Output Gate



Advantage:

1. Capturing Long-Term Dependencies
2. Sequential Information Processing
3. Less complexity and cost

Classification Report:	precision	recall	f1-score	support
Class 0	0.98	0.99	0.99	4112
Class 1	0.99	0.98	0.98	3025
accuracy			0.98	7137
macro avg	0.99	0.98	0.98	7137
weighted avg	0.98	0.98	0.98	7137

Illegal Text Detection Model Selection

Compared to machine learning algorithms, deep learning demonstrates significantly better performance.

- Among the four methods we evaluated, FastText and BERT stood out with the highest performance metrics.
- Consequently, FastText and BERT were selected as the final models for upstream filtering.

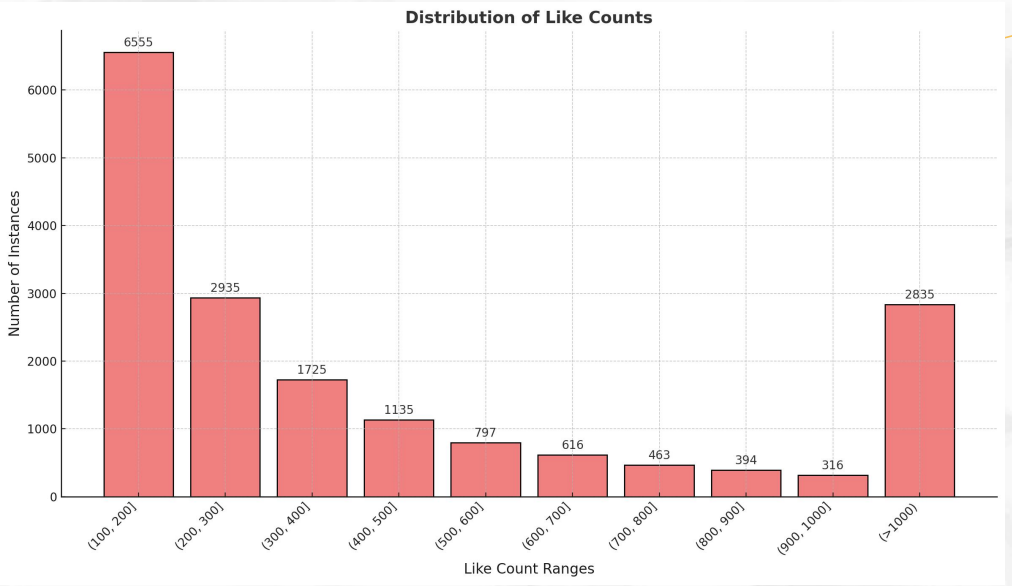
Model	Accuracy	Precision	Recall	F1-Score
FastText	0.99	0.99	0.985	0.985
Bert	0.99	0.985	0.99	0.99
LSTM	0.99	0.99	0.985	0.99
TextCNN	0.98	0.985	0.985	0.985

PART 04

High-Quality Comment Filtering

High-Quality Comments

- **Assumption**:
  Comments with high likes count are high-quality.
- **Data Source**:
  Douban Highly Liked Comments



内容	点赞数
“北京道路安全委提醒你：道路千万条，安全第一条，行车不规范，亲人两行泪”这句广播语真是又土又洗脑，能不能押点韵啊？哈哈哈~ 电影比预期要更恢弘磅礴，晨昏线过后的永夜、火种计划、让地球流浪、木星推动地球…等等大小设定，没想到中国也能拍这么大架构、大格局的科幻片了，而且是第一部，了不得。以前看国外科幻感觉离我们很远，这一次看到熟悉的北京大裤衩、上海东方明珠都变成零下89°冰天冻地的末世场景，既猎奇又唏嘘。虽然在剧情上有套路，对于这部中国文化背景下的科幻新生儿，鼓励多于挑剔。导演说美国人拍科幻是放弃地球、去挖掘新的人类居住地，而中国人是不放弃地球、守住家土的情怀…“希望是我们回家的唯一方向”	74384
王传君所有不被外人理解的坚持，都在这一刻得到了完美释放。他不是关谷神奇，他是王传君。 你看，即使依旧烂片如云，只要还有哪怕极少的人坚持，中国影视也终于还是从中生出了茁壮的根。 我不是药神，治不好这世界。但能改变一点，总归是会好的。	50970
陈凯歌可以靠它吃两辈子饭了，现在看来江郎才尽也情有可原	40058
这不是“中国版《达拉斯买家俱乐部》”，这是中国的真实事件改编的中国电影，是属于我们自己的电影，不知道就去百度一下“陆勇”。	37110
白居易和空海分别获得各自朋友圈当日微信运动冠军。	34072

Feature Extraction From Comment

Comment Data

- Unstructured Data
- Unfixed length
- Text data cannot be calculated

Converts words into vector representations **---- Word2Vec**

- **Traditional** Word2Vec generates static word embeddings by learning semantic relationships through context prediction, without adapting to different contexts.

  - Continuous Bag of Words (CBOW)
  - Skip-Gram

Word2Vec by BERT

BERT (Bidirectional Encoder Representations from Transformers)

It uses a **Transformer** architecture to process text bidirectionally, considering the full context of a word from both the left and right sides.

Key Features

1. **Contextualized Word Representations**:
   - Unlike static embeddings, BERT generates dynamic embeddings based on word context.
2. **Pre-trained and Fine-tuned**:
   - Pre-trained on large datasets and fine-tuned for specific tasks like classification, Q&A, etc.
3. **Bidirectional Understanding**:
   - Processes text both ways for deeper semantic understanding.

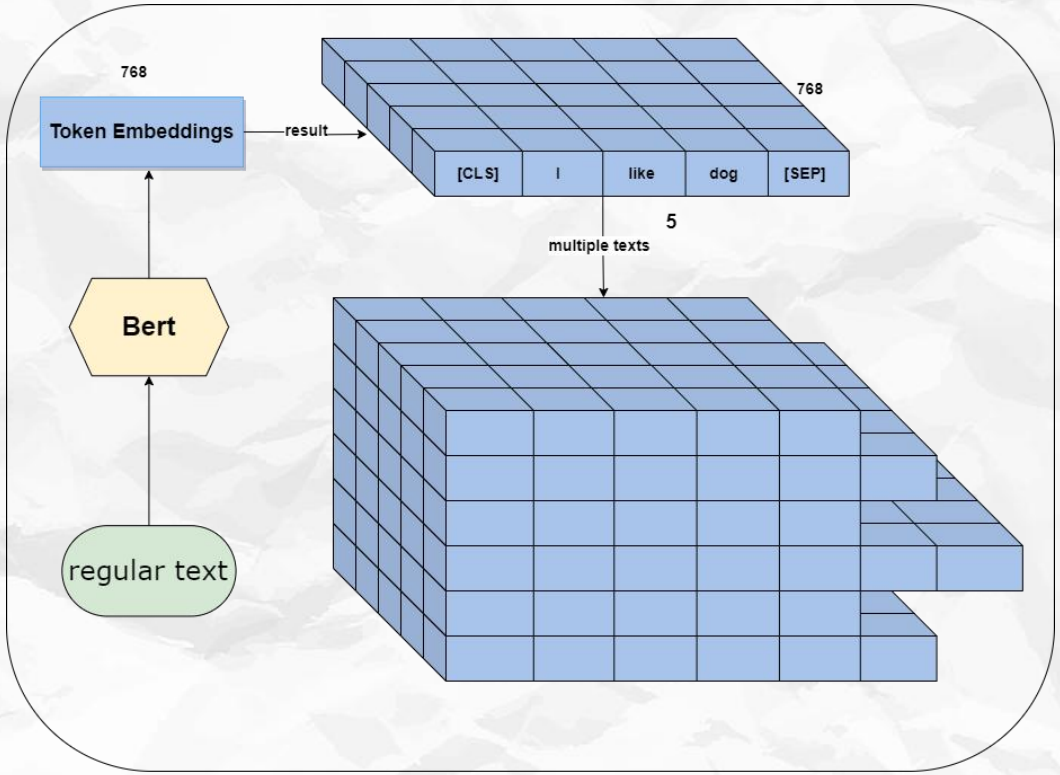
Word2Vec by BERT

BERT (Bidirectional Encoder Representations from Transformers)

It uses a **Transformer** architecture to process text bidirectionally, considering the full context of a word from both the left and right sides.

Process

- Input the comment text into BERT to get a matrix representation.



High-Quality Comment Classification

Why not just use the previous text classification model?

- **Likes ≠ Quality**:
  - While likes can indicate high-quality comments, they are not always a reliable measure.
  - For instance, **low visibility** can lead to fewer likes for otherwise valuable comments.
- **Subjectivity of Manual Screening:**
  - Manually evaluating comments is subjective and limited in scope.
  - High-quality comments gain validation when reviewed by a large and diverse audience, something manual methods can’t replicate efficiently.

High-Quality Comment Classification

How do we approach this ?

Feature Extraction with Auto-Encoders

- Identify **latent structures** in data.
- Enable discovery of subtle patterns,
  such as tone, relevance, or emotional impact, that contribute to perceived quality.

Auto-Encoder for Filtering

Auto-Encoder application in anomaly detection

- Fraud Detection in Transactions
- Network Intrusion Detection
- Manufacturing Defect Detection
- Healthcare – Detecting Abnormal ECG Signals

Auto-Encoder anomaly detection process

- **Collect Data**

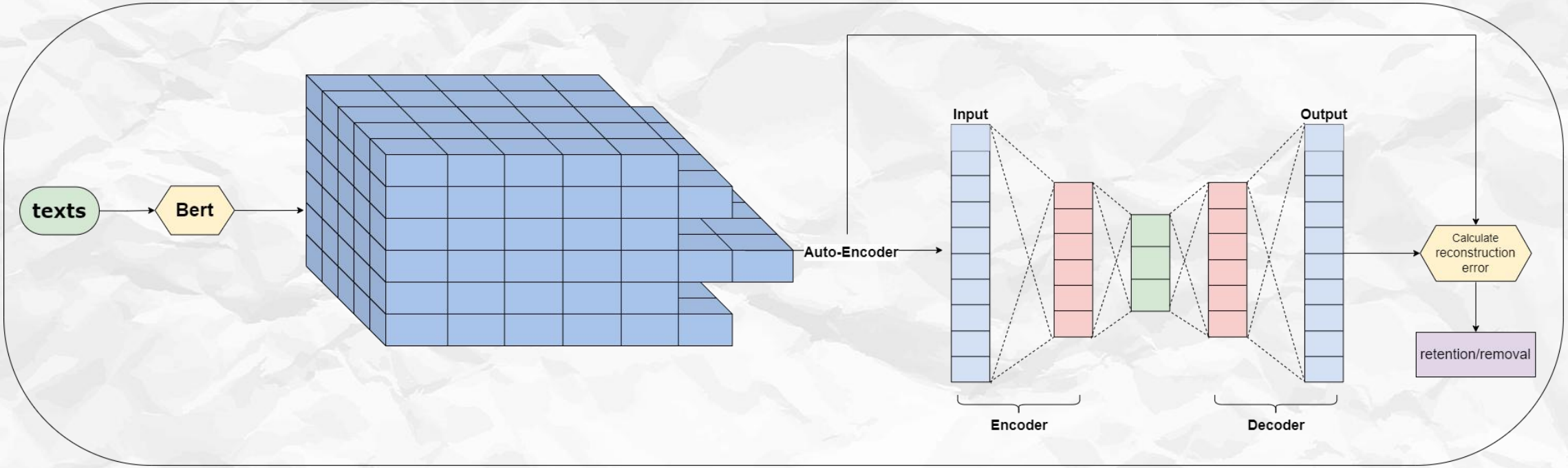
  Gather labeled or unlabeled data, ensuring that the majority represents normal patterns.

- **Encoder**: Choose an encoder to compress the input data into a lower-dimensional latent space.
- **Decoder**: Design a decoder to reconstruct the data back to its original form.
- **Train the autoencoder on normal data only.**

High-Quality Comment Classification

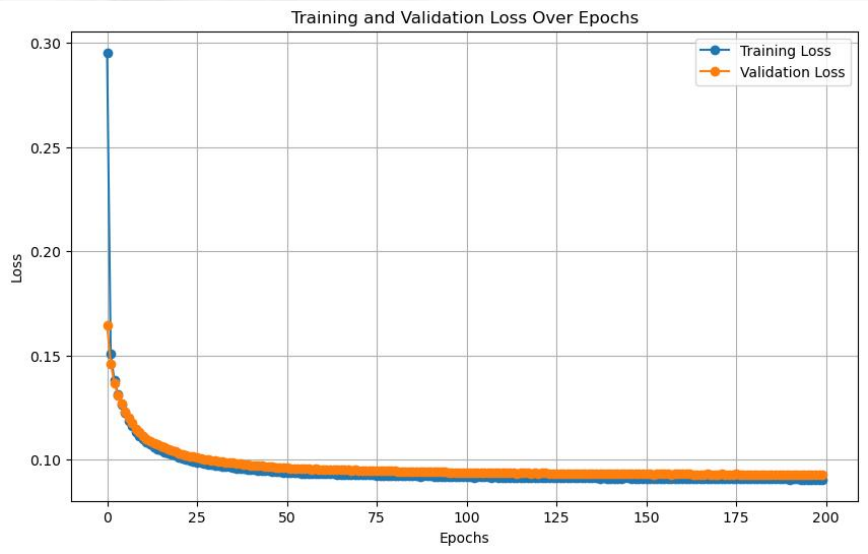
Model Architecture

- Encoder: 2 Fully connected layers and activated by Relu.
- Compressed representation with size: 32.
- Decoder: 2 Fully connected layers and activated by Relu.

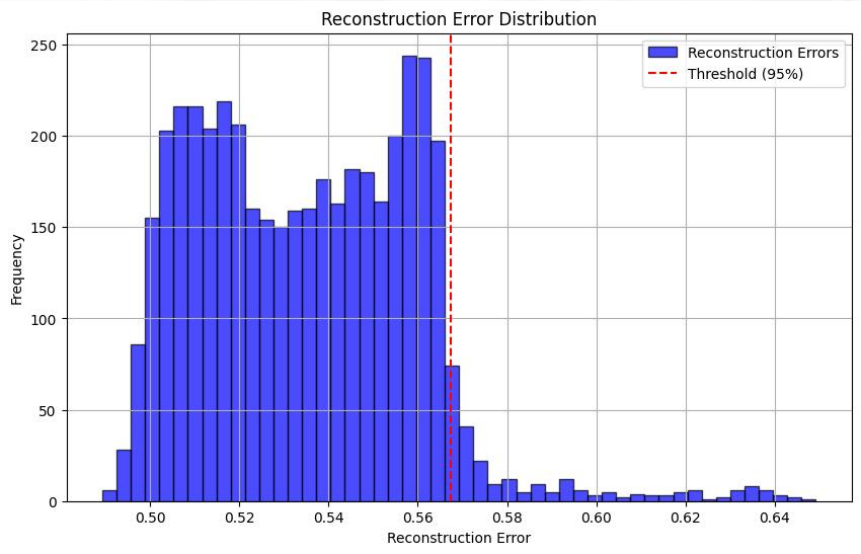


Auto-Encoder for Filtering comments

Training and Validation Loss Over Epochs



Reconstruction Error Distribution



Future Work

Validation of Effectiveness

- Since this is an **unsupervised model**, it is challenging to directly verify the effectiveness of the comment filtering process.
- **Possible Solution**: Deploy the model in a real-world application and evaluate its performance through **user feedback**.

Improving the Auto-Encoder Architecture

- The current architecture uses **fully connected layers** for both the encoder and decoder, which may be overly simplistic.

Proposed Improvements:

1. Introduce **convolutional layers** to better capture spatial or sequential relationships in text features.
2. Experiment with **Transformer-based architectures** like BERT or GPT for more sophisticated feature extraction.
3. Implement **variational auto-encoders (VAEs)** to improve generalization and uncover latent structures in the data.

Contribution

Thanks For Your Attention!