# Exploring Sentence Pair Classification Models: A Comprehensive Analysis


## Abstract

This study delves into the realm of sentence pair classification, employing state-of-the-art models to discern relationships between textual elements. Our investigation centers on the use of Random Forest (RF) and Convolutional Neural Network (CNN) architectures, supplemented by advanced transformer-based embeddings like RoBERTa. The primary objective is to find the effectiveness of these models in capturing semantic nuances within sentence pairs.


## Introduction

Understanding the relationships between sentences is a fundamental challenge in natural language processing. This study focuses on the classification of sentence pairs, a task with diverse applications ranging from information retrieval to question answering.

## Methodology

### Data Preprocessing

The primary focus was on practical preprocessing steps, including the removal of words with exceptionally low frequency throughout the entire text. Data augmentation was attempted by flipping pairs, though this technique did not yield improved results. The chosen embeddings included Count Vectorizer, TF-IDF, and a custom-trained tokenizer inspired by RoBERTa. The tokenizer has been trained with different values for vocabulary size, like 5000, **15000** and 25000.

### Model Architectures

Our exploration extended beyond conventional models, venturing into the intricacies of Convolutional Neural Networks (**CNNs**) and Support Vector Machines (**SVMs**), alongside the ensemble capabilities of Random Forest (**RF**). There was also an attempt to train a RoBERTa – like model on our corpus text from scratch. However, I didn't got satisfaction with the training metrics on the task of predicting masked tokens, and due to huge usage of GPUs for this technique, I preferred to add it to future work.

## CNN Configuration Details

The CNN architecture began with an embedding layer, serving as the foundational representation of the input sentences. Subsequently, a 1D convolutional layer, employing a kernel size of 7, facilitated the extraction of long sequential patterns. This was followed by a combination of pooling layers that contributed to dimensionality reduction while keeping essential features and residual convolutions – with a conventional kernel size of 3 - to go deeper with the features characteristics. This was ended up by a fully connected layer. The unique approach of treating word embeddings as channels in the convolutional layers enriched the model's understanding of semantic relationships.

Additionally, parameters such as the learning rate, optimizer, and linearly decreasing learning rate to zero were meticulously tuned. The weighted cross-entropy loss function was employed to address class imbalances within the dataset. Below is a list of the main tried values:

- Batch size: 8, 16, 32, **64**, 128
- Embedding size: 32, 64, **128**, 256
- Learning rate: **1e-2**, 1e-3, 1e-4

## SVM and Random Forest Configurations

In parallel, the study explored the efficacy of Support Vector Machines (SVMs) with varying C values to soften the bounds, like 1, 0.1 and **0.01**, using the Count Vectorizer for feature extraction. The Random Forest (RF) model, known for its ensemble learning prowess, was tested with different vectorization techniques, including **Count Vectorizer**, TF-IDF, and a custom-trained tokenizer inspired by RoBERTa and **with** and without **class weights**.

## Results and Analysis - Performance Metrics

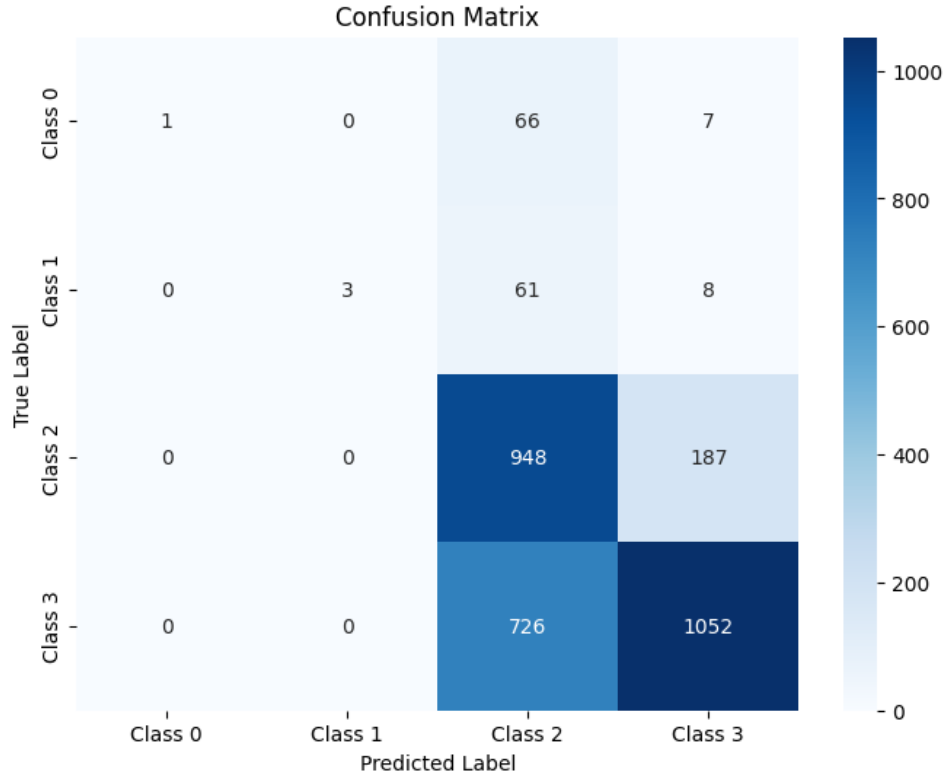
Quantitative evaluation involved precision, recall, and F1-score metrics. The RF model showed commendable metrics. Meanwhile, the CNN, demonstrated proficiency in capturing hierarchical relationships within sentences.

Below are the confusion matrixes and the classification reports for the:

- Random Forest – without class weights, Count Vectorizer encoding

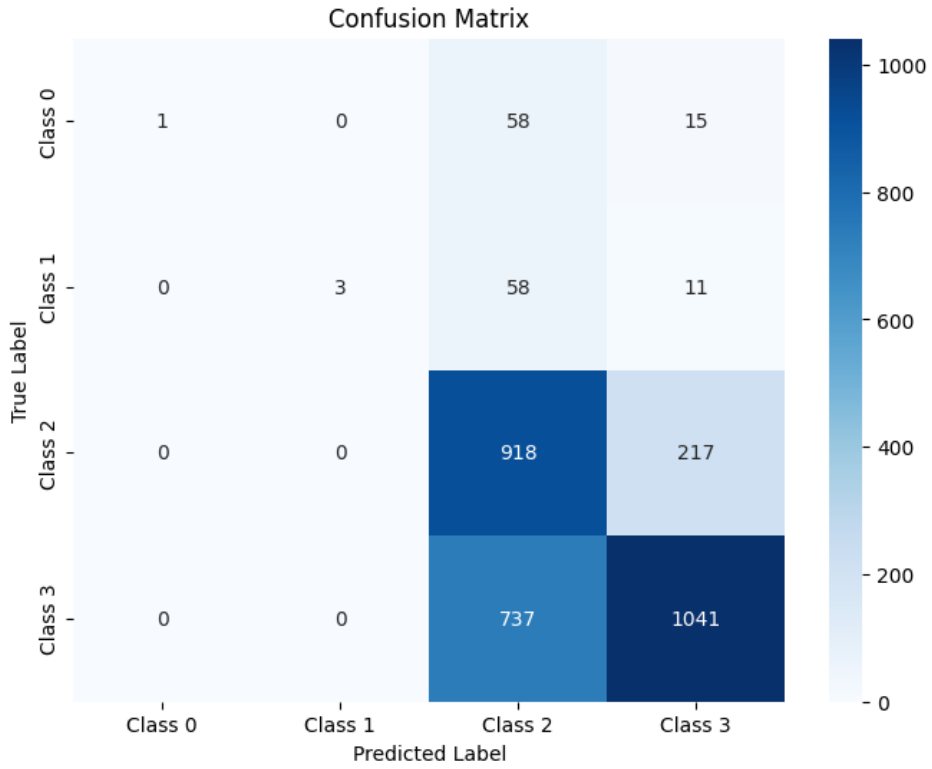
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.01 | 0.03 | 74 |

| | | | | |
|---|---|---|---|---|
| 1 | 1.00 | 0.04 | 0.08 | 72 |
| 2 | 0.53 | 0.84 | 0.65 | 1135 |
| 3 | 0.84 | 0.59 | 0.69 | 1778 |
| accuracy | | | 0.66 | 3059 |
| macro avg | 0.84 | 0.37 | 0.36 | 3059 |
| weighted avg | 0.73 | 0.66 | 0.65 | 3059 |




- Random Forest – with class weights, Tf-idf encoding

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 1.00 | 0.01 | 0.03 | 74 |
| 1 | 1.00 | 0.04 | 0.08 | 72 |
| 2 | 0.52 | 0.81 | 0.63 | 1135 |
| 3 | 0.81 | 0.59 | 0.68 | 1778 |
| accuracy | | | 0.64 | 3059 |
| macro avg | 0.83 | 0.36 | 0.35 | 3059 |
| weighted avg | 0.71 | 0.64 | 0.63 | 3059 |

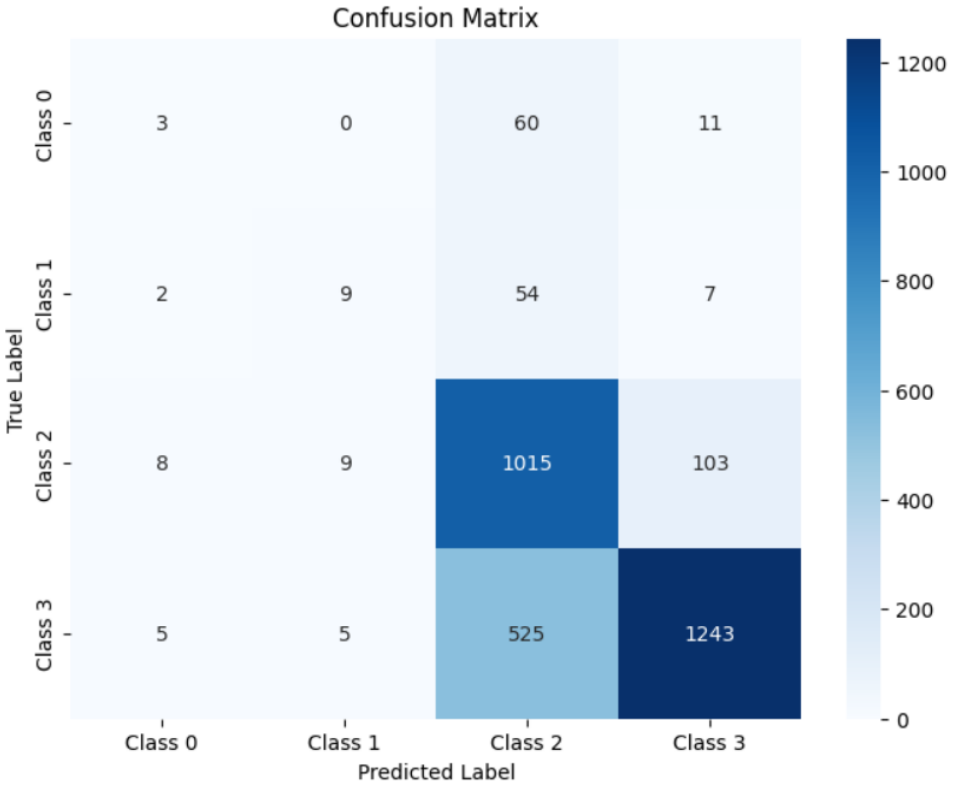



- SVM – Count Vectorizer encoding

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 74 |
| 1 | 0.00 | 0.00 | 0.00 | 72 |
| 2 | 0.51 | 0.48 | 0.49 | 1135 |
| 3 | 0.67 | 0.74 | 0.70 | 1778 |
| accuracy | | | 0.61 | 3059 |
| macro avg | 0.29 | 0.31 | 0.30 | 3059 |
| weighted avg | 0.57 | 0.61 | 0.59 | 3059 |

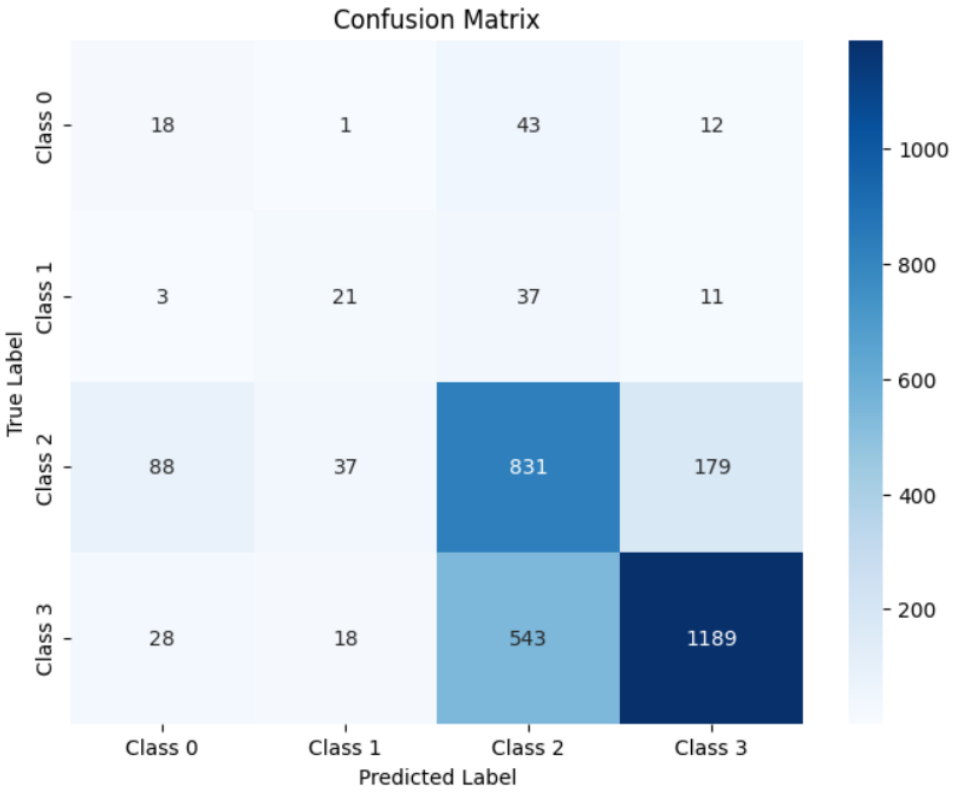
- CNN – without weighted cross entropy

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.17 | 0.04 | 0.07 | 74 |
| 1 | 0.39 | 0.12 | 0.19 | 72 |
| 2 | 0.61 | 0.89 | 0.73 | 1135 |
| 3 | 0.91 | 0.70 | 0.79 | 1778 |
| accuracy | | | 0.74 | 3059 |
| macro avg | 0.52 | 0.44 | 0.44 | 3059 |
| weighted avg | 0.77 | 0.74 | 0.74 | 3059 |




- CNN – with weighted cross entropy

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.13 | 0.24 | 0.17 | 74 |
| 1 | 0.27 | 0.29 | 0.28 | 72 |
| 2 | 0.57 | 0.73 | 0.64 | 1135 |
| 3 | 0.85 | 0.67 | 0.75 | 1778 |
| accuracy | | | 0.67 | 3059 |
| macro avg | 0.46 | 0.48 | **0.46** | 3059 |
| weighted avg | 0.72 | 0.67 | 0.69 | 3059 |




## Discussion

An intriguing observation surfaced during our analysis—the false negatives and false positives from RF and CNN models appeared to be complementary. This discovery raises the prospect of combining their predictions to achieve a more robust classification.

## Future Directions

### Ensemble Approaches

Building upon the observation of complementary errors, future research could explore ensemble techniques. Combining the strengths of RF and CNN models, possibly through weighted averages, may yield a more nuanced and accurate predictive framework.

### Attention Mechanisms and Advanced Transformers

Incorporating attention mechanisms into CNN architectures and exploring advanced transformer models beyond RoBERTa could unlock further potential for understanding intricate sentence pair relationships.

# Conclusion

In conclusion, this study provides valuable insights into the efficacy of RF and CNN models for sentence pair classification. The observed complementarity in errors between these models opens avenues for future research in ensemble approaches and advanced transformer-based architectures.