Multi Label Text Classification Using Transformers

Srujay Reddy
303207866

Samah Eltayeb
302740659

Alekya Paladugu
302845283

ABSTRACT

In multi label text classification, instances are categorised into multiple labels simultaneously. Traditional models often overlook the interdependencies between these labels. Our project introduces an enhanced model using Transformer architecture, particularly BERT, to address this gap. We utilised the Toxic Comment dataset, comprising labels such as Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. Our findings indicate that the BERT model outperforms the current state of the art MAGNET model [1], which relies on a Graph Neural Network with Attention for label dependency capture.

KEYWORDS

Multi Label Classification, Deep Learning, Transformers, MAGNET, BERT

ACM Reference Format:
Srujay Reddy, Samah Eltayeb, and Alekya Paladugu. 2023. Multi Label Text Classification Using Transformers. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Classification tasks in machine learning can be categorised into Binary Classification, Multiclass Classification, and Multi-label Classification. While Binary and Multiclass Classifications deal with mutually exclusive class assignments, Multi-label Classification presents a more complex scenario where an instance may belong to multiple classes simultaneously. Our focus is on enhancing the efficiency of Multi-label Classification models. We chose the Toxic Comment dataset for its challenging nature, encompassing 159,571 samples across six nuanced labels [2].

To approach this task, we implemented various models, starting with classical machine learning models, followed by deep learning models like BiLSTM and CNN, and then progressing to more advanced models like MAGNET (Multi-Label Text Classification using Graph Neural Network with Attention). We also explored the potential of BERT, a Transformer model, for this classification task. Our contributions include a few classical models, implementations of BiLSTM with Glove and BERT Embeddings, CNN with these embeddings, MAGNET model adaptations, and finally, the standalone BERT transformer model.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 PROBLEM FORMULATION

Multi-Label Text Classification (MLTC) poses a unique challenge in accurately assigning multiple labels to a single input sample. This complexity is compounded by the interdependencies among labels, a factor often overlooked in conventional models. Our project proposes a model that effectively captures these label dependencies, aiming to surpass the performance of existing models.

3 DESIGN OF APPROACH

3.1 System Architecture

Our project incorporates a range of classical and deep learning algorithms, each with its unique approach to the multi-label classification problem.

3.2 Classical Machine Learning Models

- Binary Relevance This traditional MLTC method treats each label as an independent classification problem, ignoring label correlations.

```
====================BR============================
              precision    recall  f1-score   support

           0       0.89      0.84      0.86      3049
           1       0.55      0.20      0.29       292
           2       0.88      0.69      0.78      1702
           3       0.71      0.11      0.19       107
           4       0.78      0.58      0.66      1564
           5       0.69      0.26      0.37       262

   micro avg       0.85      0.68      0.76      6976
   macro avg       0.75      0.45      0.53      6976
weighted avg       0.84      0.68      0.74      6976
 samples avg       0.37      0.33      0.34      6976
```

Figure 1: Classification Report for Binary Relevance.

- Classifier Chain An extension of the binary relevance method, this approach considers label dependencies while maintaining the simplicity of binary relevance.

```
====================Classifier Chain============================
              precision    recall  f1-score   support

           0       0.89      0.84      0.86      3049
           1       0.54      0.18      0.28       292
           2       0.87      0.72      0.79      1702
           3       0.69      0.10      0.18       107
           4       0.73      0.63      0.68      1564
           5       0.75      0.30      0.43       262

   micro avg       0.84      0.70      0.76      6976
   macro avg       0.74      0.46      0.53      6976
weighted avg       0.82      0.70      0.75      6976
 samples avg       0.36      0.34      0.34      6976
```

Figure 2: Classification Report for Classifier Chain.

- One Vs Rest Each label is fitted with its classifier, often combined with Logistic Regression. Its downside is the need for a separate model for each label.

==========One Vs Rest==========

	precision	recall	f1-score	support
0	0.89	0.84	0.86	3049
1	0.55	0.20	0.29	292
2	0.88	0.69	0.78	1702
3	0.71	0.11	0.19	107
4	0.78	0.58	0.66	1564
5	0.69	0.26	0.37	262
micro avg	0.85	0.68	0.76	6976
macro avg	0.75	0.45	0.53	6976
weighted avg	0.84	0.68	0.74	6976
samples avg	0.37	0.33	0.34	6976

Figure 3: Classification Report for One Vs Rest.

- Label PowerSet This approach converts the multi-label problem into a single-label classification by treating each label combination as a unique class.

==========Label PowerSet==========

	precision	recall	f1-score	support
0	0.91	0.76	0.83	3049
1	0.58	0.13	0.21	292
2	0.88	0.62	0.73	1702
3	0.83	0.05	0.09	107
4	0.78	0.55	0.64	1564
5	0.68	0.18	0.29	262
micro avg	0.87	0.62	0.72	6976
macro avg	0.78	0.38	0.46	6976
weighted avg	0.85	0.62	0.70	6976
samples avg	0.34	0.29	0.30	6976

Figure 4: Classification Report for Label PowerSet.

- Hierarchical SVM

==========Hierarchical SVM==========

	precision	recall	f1-score	support
0	0.89	0.85	0.87	3049
1	0.71	0.06	0.11	292
2	0.88	0.74	0.80	1702
3	0.63	0.11	0.19	107
4	0.76	0.63	0.69	1564
5	0.73	0.31	0.43	262
micro avg	0.86	0.71	0.78	6976
macro avg	0.77	0.45	0.52	6976
weighted avg	0.84	0.71	0.76	6976
samples avg	0.38	0.35	0.35	6976

Figure 5: Classification Report for Hierarchical SVM.

5 BiLSTM_glove 0.824693 0.077262 0.658860

	precision	recall	f1-score	support
0	0.89	0.87	0.88	3049
1	0.71	0.05	0.10	292
2	0.80	0.82	0.81	1702
3	0.00	0.00	0.00	107
4	0.69	0.74	0.72	1564
5	0.75	0.01	0.02	262
6	0.90	0.93	0.91	3245
micro avg	0.84	0.81	0.82	10221
macro avg	0.68	0.49	0.49	10221
weighted avg	0.83	0.81	0.80	10221
samples avg	0.83	0.83	0.82	10221

Figure 6: Classification Report for BiLSTM and GloVe.

3.3 BiLSTM Models

BiLSTM networks, combined with Glove and BERT embeddings, were used to obtain feature vectors for classification.

- BiLSTM with GloVe
- BiLSTM with BERT

6 BiLSTM_bert 0.827579 0.076183 0.659630

	precision	recall	f1-score	support
0	0.87	0.91	0.89	3049
1	0.64	0.15	0.25	292
2	0.81	0.83	0.82	1702
3	0.00	0.00	0.00	107
4	0.69	0.73	0.71	1564
5	0.00	0.00	0.00	262
6	0.92	0.90	0.91	3245
micro avg	0.84	0.81	0.83	10221
macro avg	0.56	0.50	0.51	10221
weighted avg	0.81	0.81	0.81	10221
samples avg	0.84	0.83	0.82	10221

Figure 7: Classification Report for BiLSTM and BERT.

3.4 CNN Models

CNNs were implemented with Glove and BERT embeddings for feature extraction.

- CNN with GloVe

7 CNN_glove 0.815251 0.080101 0.652234

	precision	recall	f1-score	support
0	0.90	0.83	0.86	3049
1	0.60	0.25	0.35	292
2	0.85	0.77	0.81	1702
3	0.00	0.00	0.00	107
4	0.72	0.66	0.69	1564
5	0.67	0.18	0.28	262
6	0.87	0.93	0.90	3245
micro avg	0.85	0.79	0.82	10221
macro avg	0.66	0.52	0.56	10221
weighted avg	0.83	0.79	0.80	10221
samples avg	0.82	0.81	0.80	10221

Figure 8: Classification Report for CNN and GloVe.

- CNN with BER

```
8          CNN_bert            0.814277      0.079947  0.660092
              precision    recall  f1-score   support

           0       0.91      0.82      0.86      3049
           1       0.68      0.19      0.29       292
           2       0.84      0.76      0.80      1702
           3       0.00      0.00      0.00       107
           4       0.72      0.65      0.69      1564
           5       0.00      0.00      0.00       262
           6       0.87      0.95      0.91      3245

   micro avg       0.85      0.78      0.81     10221
   macro avg       0.57      0.48      0.51     10221
weighted avg       0.82      0.78      0.79     10221
 samples avg       0.83      0.81      0.80     10221
```

Figure 9: Classification Report for CNN and BERT.

3.5 MAGNET (Attention-based Graph Neural Network)

Figure 10, presented below, illustrates the architecture of the MAGNET model, which is an amalgamation of two distinct models. In this setup, the Bidirectional Long Short-Term Memory (BiLSTM) network is utilised for generating feature vectors, taking word embeddings as input. Concurrently, the Graph Attention Network processes the Adjacency matrix and label vectors, outputting label-specific features. These label features are then integrated with the feature vectors produced by the BiLSTM, creating a comprehensive model that effectively captures the nuances of multi-label text classification [1].

3.6 BERT Model

We leveraged the pre-trained BERT model, adapting it to our multi-label classification context through specialised preprocessing, classification layers, and evaluation methods [3]. These modifications allowed us to effectively harness BERT's advanced language processing capabilities for our specific classification task, demonstrating its versatility in handling complex NLP challenges.

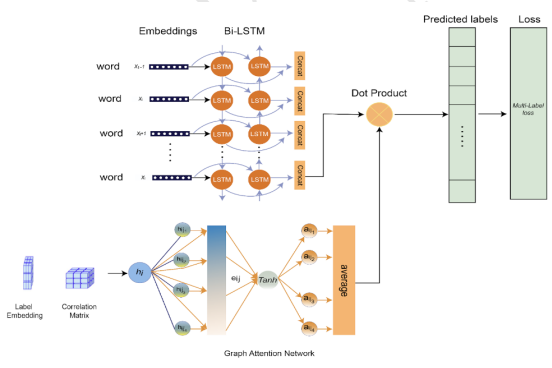

Figure 10: Architecture of MAGNET model.

```
{'model_name': 'BERT', 'micro_avg_f1_score': 0.8568846037963243,
              precision    recall  f1-score   support

           0       0.88      0.95      0.91      2447
           1       0.46      0.39      0.43       251
           2       0.84      0.87      0.85      1364
           3       0.49      0.41      0.45        68
           4       0.74      0.77      0.75      1294
           5       0.58      0.53      0.55       203
           6       0.95      0.91      0.93      2600

   micro avg       0.85      0.86      0.86      8227
   macro avg       0.70      0.69      0.70      8227
weighted avg       0.85      0.86      0.86      8227
 samples avg       0.86      0.87      0.85      8227
```

Figure 11: Classification Report for BERT Classifier.

4 EXPERIMENTAL EVALUATION

4.1 Methodology

We sourced the Toxic Comment dataset from Kaggle, comprising various labels of toxicity. The dataset was split using the train_test_split method, with 20% reserved for testing and 80% for training. Our aim was to benchmark the BERT model against the baseline MAGNET model and other classical ML algorithms. To achieve this, we followed the necessary steps described below:

- Performed Data Preprocessing by removing stop-words, making the dataset less biased.
- Implemented different models ranging from classical machine learning models, MAGNET and BERT.
- Fine-tuned BERT according to our dataset by adding a classification layer on top of the core BERT model.
- Analysed the models by implementing a classification report that contains precision, recall and F1-score.

4.2 Results

Our comprehensive evaluation of various models yielded insightful performance metrics. As seen in our results, the BERT model outshined others, achieving a micro-average F1 score of approximately 0.856, indicating its strong ability to handle multi-label classification tasks. It also showed a lower hamming loss of around 0.065, suggesting fewer misclassifications across labels, and an accuracy of approximately 0.684, the highest among all models tested.

In comparison, traditional machine learning approaches like Binary Relevance (BR), Classifier Chain, One_vs_Rest, and Label Powerset showed a range of micro-average F1 scores from about 0.759 to 0.722. These models demonstrated a decent grasp over the task but were outperformed by the deep learning approaches. Hierarchical SVM, a classical machine learning model, stood out with a micro-average F1 score of around 0.776 and an accuracy of approximately 0.672, showing its effectiveness in multi-label classification tasks.

Deep learning models, specifically those employing BiLSTM with Glove and BERT embeddings, and CNN with Glove and BERT embeddings, displayed an improvement in performance with micro-average F1 scores ranging from about 0.815 to 0.827. These models benefitted from the advanced representation capabilities of Glove and BERT embeddings.

	model_name	micro_avg_f1_score	hamming_loss	accuracy
0	BR	0.758944	0.077863	0.661787
1	Classifier_Chain	0.764628	0.077581	0.668259
2	One_vs_rest	0.759005	0.077838	0.661787
3	Label_Powerset	0.722120	0.085362	0.653621
4	Hierarchical_SVM	0.775363	0.073523	0.671957
5	BiLSTM_glove	0.824693	0.077262	0.658860
6	BiLSTM_bert	0.827579	0.076183	0.659630
7	CNN_glove	0.815251	0.080101	0.652234
8	CNN_bert	0.814277	0.079947	0.660092
9	MAGNET_Cooccurance	0.820034	0.078935	0.640216
10	MAGNET_xavier	0.818782	0.079177	0.644838
11	MAGNET_random	0.815133	0.081532	0.630200
12	BERT	0.856885	0.065247	0.684615

Figure 12: Results for all implemented models.

The MAGNET model variants, Cooccurrence, Xavier, and Random, yielded micro-average F1 scores between approximately 0.818 and 0.820. While these scores were competitive, they did not reach the high benchmark set by the BERT model. The results underscore BERT's exceptional capability in text classification, likely due to its deep bidirectional nature, allowing it to contextually understand text better than other models tested.

These findings indicate that while traditional and some deep learning models hold merit in multi-label text classification, the advanced architecture of BERT provides a significant edge, potentially setting a new standard for such tasks.

5 RELATED WORK

In multi-label text classification, recent advancements have centred around capturing label interdependencies and text representation. The MAGNET model, introduced by Pal et al. (2020) [1], leveraged attention-based graph neural networks to map label relationships. The introduction of BERT by Devlin et al. (2019) [3] provided a breakthrough in language representation, utilising deep bidirectional training. Our project extends these works by empirically validating BERT's effectiveness over traditional GNNs on the Toxic Comment dataset [2], illustrating its potential for nuanced text classification tasks.

6 CONCLUSION

This project demonstrates the robustness of the BERT model in handling complex multi-label text classification tasks, outperforming traditional models like MAGNET and other classical models. This advancement could have significant implications in various NLP applications requiring nuanced label categorizations. Future work could explore further optimizations of the BERT model and its application to other complex text classification

7 WORK DIVISION

All three of us contributed equally and we all worked on all the components of the project.

8 LEARNING EXPERIENCE

Our project was a deep dive into the practicalities of machine learning, providing our team with a rich learning experience in multi-label classification. We tackled the complexities of preprocessing text data, fine-tuning BERT and various other models, and critically analysing results using robust evaluation metrics like F1 score and hamming loss. In navigating the nuances of Transformer models, especially BERT, we enhanced our collective understanding of sequential data processing and contextual analysis in NLP.

As a team, we strengthened our ability to communicate complex ideas and collaboratively solve challenging problems. The project bridged theoretical knowledge from our studies with practical application, underlining the dynamic and ever-evolving nature of machine learning. This experience has solidified our commitment to continuous learning and adaptation in the field of AI.

9 REFERENCES

(1) A. Pal, M. Selvakumar, and M. Sankarasubbu, "MAGNET: Multi-Label Text Classification using Attention-based Graph Neural Network," Proceedings of the 12th International Conference on Agents and Artificial Intelligence, 2020.
(2) cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, Will Cukierski. (2017). Toxic Comment Classification Challenge. Kaggle.
(3) Jacob Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.