**CS-412 Report**

The project leveraged pre-trained embedding models, advanced neural networks for classification, and traditional machine learning algorithms for like count prediction to evaluate the effectiveness of different approaches. The dataset included profiles, posts, and associated metadata for various Instagram influencers. First captions with fewer than 10 characters or those are considered as meaningless, such as those containing only emojis or non-alphanumeric content. This filtering was critical because captions lacking linguistic content provided no value for classification or prediction tasks. Furthermore, user level features like average like counts were calculated to capture contextual information about user engagement, which proved valuable in predicting post popularity.

One of the primary challenges encountered was the quality of the data itself. Many captions did not reflect anything meaningful about the influencer's identity or content. Some captions consisted entirely of emojis or nonsensical phrases, making it difficult to associate them with a specific influencer category. This lack of context posed significant challenges for both classification and prediction tasks, as models rely on meaningful textual features to draw inferences. The presence of such data noise highlighted the importance of robust filtering and pre-processing methods to ensure that only valuable captions contribute to the analysis.

The complexity of the task further added to the challenges. Influencer classification required models to understand the subtle semantics of captions and their alignment with specific categories, making it difficult for basic neural networks to perform well. The size of the dataset further highlighted the limitations of traditional machine learning algorithms like logistic regression and random forests, which failed to capture the nuanced patterns present in the data. Also, SVM is useless because of its time complexity, for this data the wait is very long. This underscored the necessity for advanced deep learning models capable of leveraging the rich embeddings provided by the pre-trained BAAI/bge-m3 model retrieved from huggingface.

The like count prediction task, however, relied on a **random forest regressor** instead of a neural network. This choice was driven by several factors. First, the task involved predicting numerical values, where structured features such as caption length, number of hashtags, media type, and user level average likes played a significant role. Random forests are well suited for structured data due to their ability to capture non linear relationships and feature interactions without extensive feature scaling or transformation. Second, while neural networks excel in high dimensional feature spaces, such as those involving embeddings, the prediction task benefited from interpretable, tabular features. This made random forests a more practical and computationally efficient choice. Lastly, the size of the dataset and the nature of the features did not necessitate the complexity of a neural network, allowing the random forest regressor to achieve satisfactory performance.

The use of pre-trained models demonstrated clear advantages over traditional approaches like TF-IDF or simple word embeddings. Unlike traditional methods that rely on surface level word frequencies and co-occurrence statistics, pre-trained models such as BAAI/bge-m3 can encode the semantic meaning of captions. These models effectively capture the contextual relationships between words, allowing for a deeper understanding of the text. For example, a caption about "fitness tips" might be semantically linked to the "health" category, even if the exact words differ across captions. This semantic encoding proved critical for influencer classification, as it allowed the model to infer meaning beyond simple keyword matches. As a result, the embeddings generated by the pre-trained model provided a strong foundation for downstream tasks, outperforming traditional methods in both accuracy and robustness.

The analysis revealed that traditional machine learning models underperformed significantly in classification tasks. Logistic regression and random forest models lacked the capacity to utilize the semantic information encoded in the embeddings, leading to suboptimal accuracy in influencer classification. On the other hand, neural networks, particularly an advanced multi-layer perceptron (MLP), showed better results. The use of batch normalization, dropout, and activation functions such as GELU improved the model's generalization to the data. However, even the advanced architecture struggled due to the nuanced nature of the task, suggesting the need for more sophisticated solutions, such as fine tuning the pre-trained embeddings or employing transformer based architectures.

The role of pre-trained embeddings was pivotal in improving performance, as they effectively captured the semantic information of captions. Despite this, the task required further optimization to fully utilize the embeddings' potential. For predicting like counts, the random forest regressor effectively leveraged features such as caption length, number of hashtags, and average user likes, highlighting its strength in handling structured tabular data.

The findings also highlighted key insights regarding influencer classification. Many captions lacked sufficient context to determine the influencer category accurately, emphasizing the importance of combining captions with additional metadata, such as hashtags or media type, for better predictions. Hard voting across multiple captions per user helped mitigate noise from individual meaningless captions and improved user level predictions. However, this approach further emphasized the need for robust filtering methods to ensure that only meaningful captions contribute to the classification process.

In conclusion, neural networks proved to be more effective than traditional machine learning models for classification tasks due to the scale and complexity of the dataset. Pre-trained models showcased their effectiveness by capturing semantic relationships in captions, far surpassing the limitations of traditional methods like TF-IDF. For like count prediction, the choice of a random forest regressor demonstrated its effectiveness in leveraging structured features and handling tabular data efficiently. The noisy and inconsistent nature of the data highlighted the need for improved caption filtering and feature extraction processes. Advanced architectures, such as transformers or domain-specific embeddings, are likely to further improve performance in classification tasks. Addressing imbalanced category distributions with techniques like weighted loss functions or oversampling and enriching the feature space with visual content and timestamp-based engagement trends can provide additional improvements. Future work could explore multi-modal approaches combining text, image, and metadata features to capture a holistic view of influencers and their content. Overall, this analysis demonstrated the strengths and limitations of different approaches and the potential of advanced AI models in tackling complex tasks like Instagram influencer classification and engagement prediction.