INTERPRETING USER SENTIMENT IN INSTAGRAM POSTS: A MULTIMODAL DEEP LEARNING APPROACH

Charlotte Vedrines
Student# 1007747253
charlotte.vedrines@mail.utoronto.ca

Zeina Shaltout
Student# 1008039187
zeina.shaltout@mail.utoronto.ca

Roza Cicek
Student# 1008062025
roza.cicek@mail.utoronto.ca

ABSTRACT

In the age of social media, Instagram serves as a prominent platform for users to express themselves through visual and textual content. However, deciphering the sentiments behind these posts, often laden with irony and cultural nuances, poses a significant challenge due to the intricate interplay between images and text. Being able to accurately classify the sentiment of an Instagram post would provide valuable insights for businesses seeking to gauge customer reception of their brand, or for sensitive users who wish to filter out negative posts from their feed. This report introduces a multimodal deep learning approach aimed at interpreting user sentiment in Instagram posts by analyzing both visual and textual elements. Our proposed model integrates two parallel deep learning networks that process images along with their associated caption and comments. The conclusion of the post's overall sentiment would be deduced from a trainable weighted average of the two model predictions. Leveraging our collective expertise and the methodologies outlined in this proposal, we aim to develop a robust sentiment analysis model that enhances user experience and engagement on Instagram across both commercial and personal spheres.

—-Total Pages: 9

1 INTRODUCTION

"A picture is worth a thousand words" is a sentiment likely echoed by Instagram's over 2 billion monthly users Since its launch in 2010, the image-based social media platform has attracted a wide variety of users, ranging from personal profiles to accounts run by brands and businesses (Instagram, 2010). The platform presents a unique challenge for natural language processing (NLP): accurately interpreting the nuanced relationship between images and texts in posts, which often contain irony or cultural references. A deep learning model is well-suited for such a task as it is capable of learning the patterns that form between image-text pairs over large datasets. Companies could gauge brand perception based on comments under their posts and businesses could assess the effectiveness of their campaigns by measuring user sentiment to their content. Furthermore, regular users could filter out content that is be harmful to their mental health, as well as monitor the sentiment of their own posts to see how their content impacts others.

1.1 PROJECT DESCRIPTION

1 visualizes the pipeline of the architecture, featuring a dual deep learning model that processes two inputs from Instagram posts: an image and its caption, through two distinct models. The sentiment class of the post is determined by a trainable weighted average, acting as an attention mechanism.

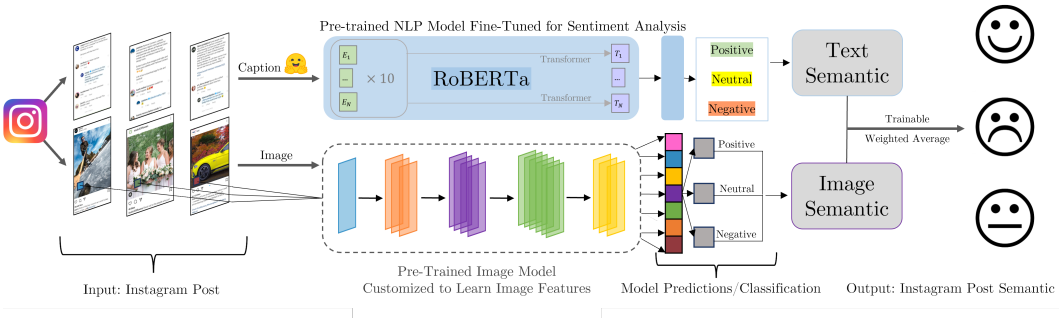


Figure 1: High-level illustration of multimodal deep learning pipeline.

2 Background and Related Work

This section consolidates findings from five research papers, each contributing insights into the development of sentiment analysis through multimodal data integration and processing techniques.

The first study introduces a model involving Multimodal Learning Fusion Convolution (MLFC) Wang et al. (2023). This approach supports our strategy of concatenating CNN and Recurrent Neural Network (RNN) outputs for sentiment classification, emphasizing the value of multimodal feature integration.

Another multimodal sentiment analysis is a fusion strategy with unimodal feature extraction, context-aware feature extraction with RNN, and hierarchical fusion, highlighting the significance of context in sentiment analysis (Majumder et al., 2018). This underlines the benefits of incorporating context-aware features to refine sentiment interpretation.

A multi-head attention mechanism to amalgamate features from audio, visual, and textual data, suggesting a method for integrating CNN and RNN outputs (Xi et al., 2020). This approach presents a direction for enhancing accuracy through attention mechanisms.

OpenAI's CLIP model introduces a unified approach to image and text interpretation, employing contrastive learning for zero-shot learning capabilities (Pinecone, 2021). This model, which leverages a modified ResNet and a Transformer (Palucha, 2024), demonstrates the power of understanding visual and textual contexts in a shared space, potentially revolutionizing sentiment classification strategies by assessing nuances in multimodal contexts (Radford et al., 2021).

These researchers use transfer learning models for image-based sentiment analysis in social media like (MathWorks, 2024), (TensorFlow, 2024), and (Keras, 2024), suggesting that pretrained models are most effective for our project's image processing component (Chandrasekaran et al., 2022).

3 Data Processing

Because our project uses three models, three labelled datasets are required to train our pipeline. A fourth dataset was collected from Instagram and labelled by our team to evaluate our final pipeline.

To assess the performance and generalisation ability of our models our data underwent a 80-20 split for the training and validation set.

Ultimately three sentiments will be output, class 0 is negative, 1 is neutral and 2 is positive.

3.1 Image Dataset

The image dataset for sentiment analysis used was the Crowdflower Sentiment Polarity dataset (CrowdFlower, 2016) with 389, 250, and 2022 images per class. To build a larger balanced dataset, the data was augmented using random rotations (up to 40 degrees), shifts in width and height (up to 20%), shear transformations (up to 20%), and zoom adjustments (by 20%); resulting in 1 400 images per sentiment class.

Preprocessing for CNN input involved resizing images to 256x256 pixels, making all the images RBG and normalizing them based on per-channel mean and standard deviation derived from the training dataset, covering the red, green, and blue channels. This approach follows practices from the high-performing model AlexNet (Delorme, 2021).

3.2 Text Processing

The language dataset used is a twitter open source dataset (Hussein, 2021), collected in 2021, contains 162,980 tweets classified, only 8000 where used. The first preprocessing step was removing NaN instances, followed by tokenization to segment sentences into individual words or tokens. Then lemmatization was applied to each token (for Geeks, 2024).

3.3 Image and Associated Caption Dataset

The dataset obtained from Flickr (Borth et al., 2013), which is used for the image dataset, also includes captions for each image. These captions will be used to train the Weighted Average NN (WANN), which is designed to determine the optimal weighting for the outputs from the RNN and CNN. A .csv file contains the captions, along with their associated sentiment scores and IDs. These IDs correspond to the images in the image folder, where the title of each image is its ID.

3.4 Custom Instagram Image and Caption Test Dataset

To properly assess the performance of our Instagram sentiment analysis pipeline, we wil test the model on a dataset collected from Instagram. Because there is no open source dataset Instagram posts (including image and text) labelled with their sentiment our team collected our own test set. Each team member collected 60 images from Instagram along with their associated captions. Each image is paired to a caption with an ID. In total, 240 images and corresponding captions were collected.

4 Architecture

4.1 Image Model: Fine-Tuned ResNet-50

4.1.1 Description and Architecture

The chosen pre-trained model is ResNet-50, a 50-layer neural network composed of 3-layer bottleneck blocks (He et al., 2015).

The weights in the model's convolutional layers are initially frozen so that only the last fully connected layer can be trained. A custom classifier is added and is devised of a linear layer of shape (num_features, 256), ReLU activation function, dropout with 50% chance of dropping, and three output classes: positive, negative, and neutral.

4.1.2 Parameters and Training Setup

A batch size of 32, a cross-entropy loss function and stochastic gradient descent is employed for optimization during training with a learning rate of 0.0005 and momentum of 0.9. The model is trained over 150 epochs using the training and validation datasets. 2 visualizes the architecture.

4.2 Text Model: Fine-Tuned RoBERTa

4.2.1 Description and Architecture

The pre-trained text model selected for this task is cardiffnlp/twitter-roberta-base-sentiment (Barbieri et al., 2020), a variant of the RoBERTa model optimized for understanding the sentiment conveyed in tweets.

To finetune the model, the twitter dataset described in section 3.2 was preprocessed. The default configuration is modified to include a custom classification layer that replaces the original fully connected output layer to ensure three sentiment classifications as integers.

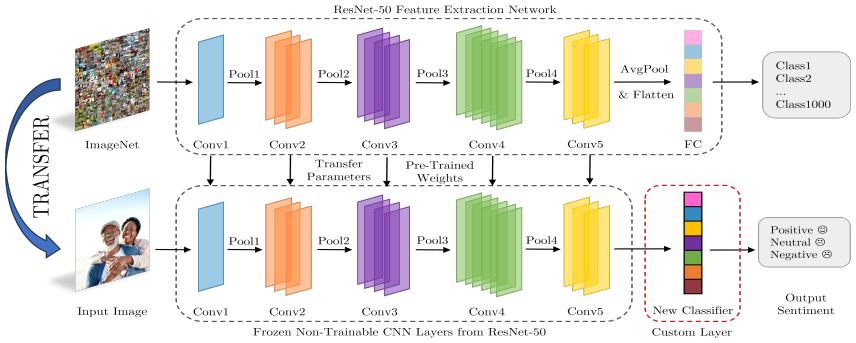


Figure 2: Architecture of image sentiment analysis CNN with transfer learning.

4.2.2 Parameters and Training Setup

The parameters are a small batch size of 8, cross-entropy for the loss function, an Adam optimizer with an initial learning rate of 0.0005, including a 500-step warm-up to stabilize early training. To avoid overfitting, we incorporate a weight decay of 0.01. There where 25 epochs of training, with early stopping and calculation of accuracy, training, validation loss, and F1 scores at each epoch's end. The best model weight where saved for subsequent the WANN training. The document includes a visual depiction of the architecture for further clarification.

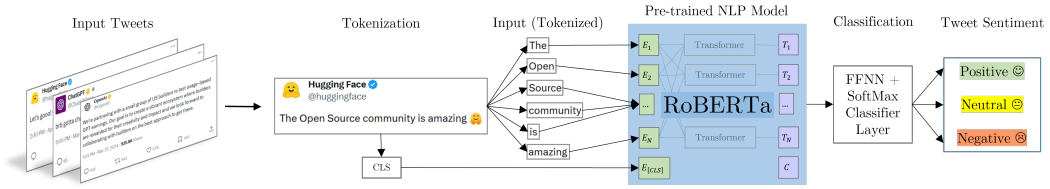


Figure 3: Diagram of the LSTM RNN architecture.

4.3 Weighted Average Neural Network (WANN)

4.3.1 Description and Architecture

The trainable weighted average is designed to learn how to optimally combine the predictions from the text and image model to make a final prediction. The goal is for the weighted average to learn which model's predictions are more reliable in what context.

The model's architecture is a Sequential model with three fully connected layers, populated with dropout and batch normalization layers in between each dense layer to avoid overfitting and enhance generalizability. The input dense layer has 64 neurons, a ReLU activation function, and regularization function. The intermediate Dense layer has 32 input neurons. The last layer uses a SoftMax activation function to output the class probability distribution of the input Instagram post.

4.3.2 Model Parameters and Training Setup

The model utilizes Adam optimizer with a fine tuned learning rate of 0.001 and the categorical cross entropy function. The model is trained with 50 epochs with early stopping engaged after 5 epochs.

5 Baseline Model

5.1 Image Baseline: Random Forest

5.1.1 Description and Comparison Strategy

For the image sentiment analysis, a Random Forest (RF) Classifier was chosen due to its efficiency in handling multi-class classification tasks (Breiman, 2001).To facilitate fair comparison between the primary model and its baseline, both were trained and evaluated on identical datasets and metrics.

The preprocessing included converting them to RGB, resizing to a standard square size, normalizing pixel values, and converting from 2D arrays into 1D arrays (V, 2023). The dataset was divided using an 80-20 split for training and testing.

The model was evaluated on accuracy, F1-score, and a confusion matrix. These metrics will be used to compare the RF Classifier's performance against the neural network on the test set.

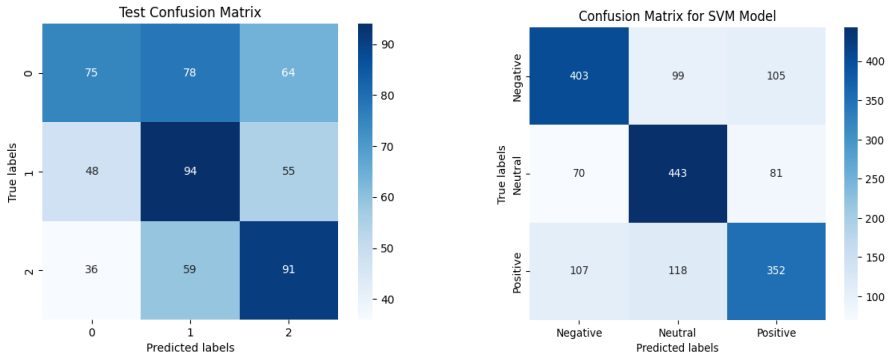
5.1.2 Implementation and Configuration

Regarding hyperparameter tuning, 5-fold cross validation was performed on RF's two main parameters: 'n_estimator', which is the number of trees in the forest and 'max_depth', the maximum depth of a tree. The chosen parameters are a value 400 for the 'n_estimator' and a 'max_depth' of 10.

5.1.3 Results

The accuracy of the model on the test dataset is 43.33%. The F1 scores for individual classes are: 0.40 for negative, 0.44 for neutral, and 0.46 for positive sentiments. These results suggest a relatively balanced performance across classes, with a slight advantage in recognizing positive sentiments.

The confusion matrix in 4a exhibits the model's tendency to confuse negative sentiments with neutral ones, as evidenced by the 78 instances incorrectly classified. Similarly, neutral sentiments are often mistaken for positive ones, with 55 instances misclassified. Positive sentiments are confused with negative ones 36 times and with neutral ones 59 times, indicating a challenge for the model to distinguish between nuanced sentiment expressions.



(a) Confusion matrix for RF Classifier

(b) Confusion Matrix for SVM

Figure 4: Confusion matrices of baseline models (0 is negative, 1 is neutral, 2 is positive)

5.2 Text Baseline: Support Vector Machine (SVM)

5.2.1 Description and Comparison Strategy

For the text sentiment analysis, a SVM was chosen. To fairly compare the primary model with the SVM, the same dataset, preprocessing methods, and evaluation metrics, including accuracy, precision, recall, F1-scores, and a confusion matrix are employed. These metrics will serve as benchmarks to evaluate the primary model's performance on the test set.

5.2.2 Implementation and Configuration

TfidfVectorizer was used to highlight important features through term frequency and inverse document frequency (Buitinck et al., 2024). Hyperparameter tuning involved a grid search that identified an optimal C value of 1, optimizing the model's generalization and mitigating overfitting risks.

5.2.3 Results

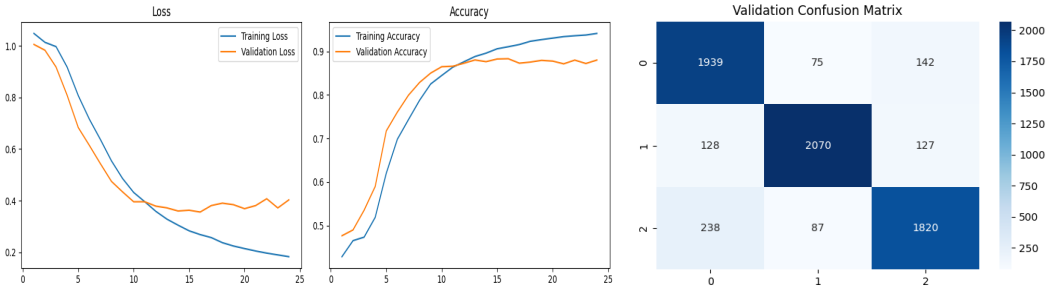
The SVM baseline achieved 77.34% accuracy, with nuanced performance across sentiment categories: it showed reliable identification of negative sentiments (precision: 0.79, recall: 0.76), better capture of neutral sentiments (precision: 0.74, recall: 0.87), and indicated room for improvement in positive sentiment detection (precision: 0.81, recall: 0.69). The confusion matrix in 5b revealed the model's strength in recognizing neutral sentiments with 2403 correct classifications, but also highlighted its difficulty in differentiating negative from positive sentiments, with 363 positive instances incorrectly labeled as negative and 497 negative instances misclassified as neutral.

6 Quantitative Results

6.1 Fine-Tuned RoBERTa Results

By epoch 24, just before early stopping was triggered, the model recorded a validation accuracy of 88.03%, and corresponding F1 scores of 0.88 for validation, indicating high proficiency in understanding and categorizing sentiments. The high F1 score on the training set indicating a robust ability to balance precision and recall across all classes. 5a displays a consistent decline of the training loss, showcasing the model's growing proficiency. Post-epoch 10 the model exhibits overfitting when the training curve surpasses the validation curve.

The confusion matrix in 5 displays a significant concentration of correct predictions along the diagonal (1939, 2070, and 1820 for negative, neutral, and positive sentiments, respectively) highlighting the model's accuracy. The most common misclassification by the model is between positive and negative sentiments (238 misclassifications from positive to negative).



(a) Loss & Accuracy Curves for training & validation of RNN (b) Confusion Matrix for SVM Base-model

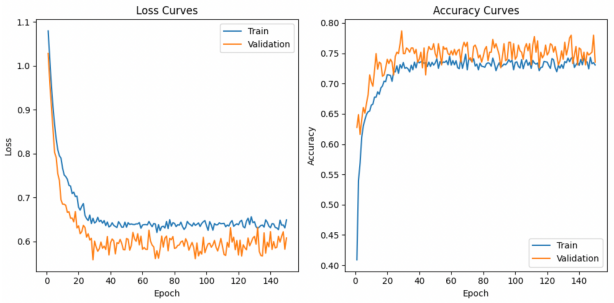
Figure 5: Results of the RoBERTa model

6.2 Fine-Tuned ResNet-50 Results

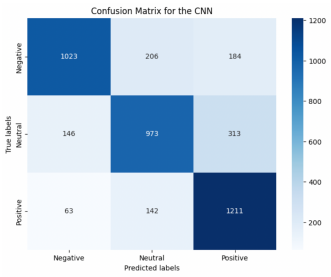
On the 150th epoch, the CNN output a maximum validation accuracy of 78.69%, indicating a good performance in its task and significant improvement from the CNN trained from scratch. As seen on the figure 6a the loss curve steadily decreases, and the accuracy curve steadily increases, indicating the model is improving, slightly overfitting in the last epochs.

The confusion matrix in 6b demonstrates that the CNN is especially proficient in correctly classifying positive sentiments.

The most significant challenge lies with neutral images where it achieved 973 correct predictions but also misclassified 146 as negative and 313 as positive, suggesting the model struggles to distinguish neutral features as effectively as it does with the more distinct positive and negative ones.



(a) Loss & Accuracy Curves for Training & Validation of CNN



(b) Confusion Matrix for CNN model

Figure 6: Results of the ResNet-50 model

6.3 WANN Results

The Flickr dataset described in section 3.3 was used to train the WANN and produce the provided confusion matrices for the model in 7. According to the confusion matrix, the WANN improves upon the RNN's and CNN's predictions, especially for the negative class; making 334 correct predictions for class 0, significantly higher than the RNN's 171 and CNN's 70, suggesting the WANN is effectively the CNN for negative sentiment, possibly giving it more weight. For the positive class, the WANN outperforms the other models, capitalizing on the RNN's ability to detect positive sentiment while incorporating the CNN's context to reduce false positives. Finally, the neutral class is where the WANN improves upon the RNN but not as much over the CNN.

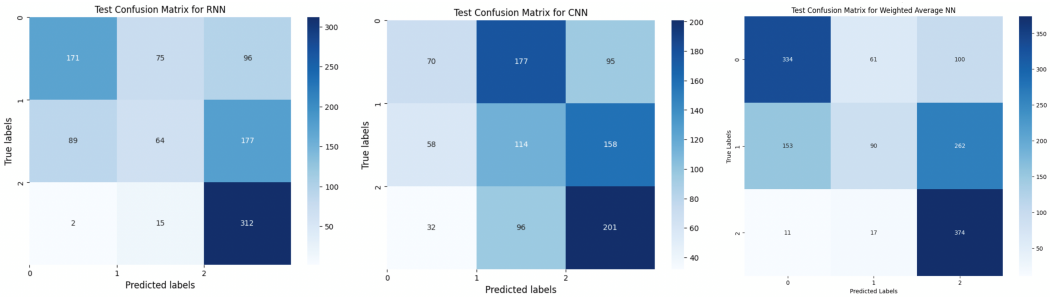


Figure 7: Confusion Matrices for WANN

7 Qualitative Results

From section 6, it is clear that the WANN finds difficulty in classifying neutral sentiment instances the most. 1 gives us an insight on some of WANN's misclassified neutral instances.

The first example in table 1 has positive connotations in both the caption and image, which could have been classified as positive by a human, exhibiting the ambiguity of neutral data points. The two other examples in table 1 suggest that the text model may be assigning sentiment based on the presence of certain keywords that are typically associated with positive or negative sentiments, rather than understanding their context.

On the other hand, 2 helps us visualise the model's success through a case correctly classified by WANN; even with misclassified and contradicting RNN and CNN outputs.

8 Evaluation of Model on New Data

The custom Instagram test dataset described in section 3.4 was used to evaluate the model. As shown in 3, the WANN, by integrating image and text model strengths, achieves a 57% accuracy and outperforms individual models, indicating the effectiveness of an ensemble approach for this problem.

Image			
Caption	There were two young deer on this very quiet stretch of road	Rio Chillar in previous years a dry river	dry leaves in vytina park in Peloponnisos Greece
CNN	Predicted Label: 2	Predicted Label: 2	Predicted Label: 2
RNN	Predicted Label: 2	Predicted Label: 0	Predicted Label: 0
WANN	Predicted Label: 2	Predicted Label: 0	Predicted Label: 0
Truth	Actual Label: 1	Actual Label: 1	Actual Label: 1

Table 1: Visualisation of neutral instances misclassified by WANN

Image	**Caption**	**Predicted Label CNN**	**Predicted Label RNN**	**WANN Predic-tion**	**Actual Label**
	Orchard Central potpurri dry dried leaves arrangement brown	2	0	1	1

Table 2: Correctly classified neutral instance by the WANN

The RNN shows exceptional recall for class 2, but its precision is lower than the WANN, indicating that the WANN is able to learn to detect a large number of positive instances from the RNN while maintaining a lower number of false positives for class 2. The CNN, while having the lowest overall performance, contributes to the WANN's ability to better identify class 0 and 1; reinforced by the increase in precision for WANN.

9 Discussion

As seen in the qualitative section, WANN can discern patterns in the outputs of the image and text model; leading to correct conclusions. Table 4 compares the number of WANN's prediction for each class, given output combinations from the image and text models. The last column calculates the average accuracy of the WANN for a combination.

WANN's performance fluctuates when CNN and RNN predictions conflict; accuracy drops to 47% with CNN negative and RNN positive disagreements, often siding with RNN's positive outlook. Neutral predictions present a challenge, with WANN aligning with both CNN and RNN's neutral verdicts in 9 out of 373 instances, achieving 60% accuracy. This indicates WANN's capacity to somewhat rectify CNN and RNN's biases or mistakes, yet its reliability in neutral sentiment classification remains comparatively low, indicating the challenges inherent in neutral sentiment analysis by CNN and RNN.

10 Ethical Considerations

Our sentiment analysis system's reliance on vast amounts of visual and textual social media data raises significant privacy and ethical concerns. The potential use of such a model by platforms like Instagram to analyze user sentiments and tailor content or advertisements could lead to privacy invasions and emotional manipulation. It's crucial that users are informed about data collection practices and have the option to opt out to protect their privacy.

	Accuracy	Precision			Recall			F1 Score		
Class		0	1	2	0	1	2	0	1	2
CNN	40%	47%	33%	43%	23%	35%	63%	31%	34%	51%
RNN	53%	63%	41%	51%	52%	17%	93%	57%	24%	66%
WANN	57%	67%	54%	51%	67%	18%	93%	57%	53%	52%

Table 3: Comparison of evaluation metrics of the three models in the pipeline. Class 0 is negative, class 1 is neutral, and class 2 is positive.

CNN Pred.	RNN Pred.	WANN Classification per Class			WANN Acc.
		0	**1**	**2**	
0	**0**	384	0	0	72%
0	**1**	185	0	0	59%
0	**2**	21	201	346	47%
1	**0**	817	4	0	72%
1	**1**	364	9	0	60%
1	**2**	22	419	943	45%
2	**0**	381	287	0	53%
2	**1**	35	404	0	48%
2	**2**	0	60	2127	61%

Table 4: Comparison of the WANN's predictions and accuracy given different combinations of the image and text output

Additionally, sentiment analysis models, particularly those using NLP to interpret image-text content, may inaccurately capture the nuances of Instagram posts, reflecting biases present in the training data. This can perpetuate discrimination and stereotypes, affecting certain demographics unfairly. To combat this, a diverse data collection strategy, bias detection and correction mechanisms, and transparency about the algorithm's workings are essential to ensure fairness, build trust, and maintain accountability with users.

11 Project Difficulty and Quality

The project faced several challenges, primarily due to the need for multiple datasets to train and test our complex pipeline of independent models. Finding a dataset with paired images and text that mirror Instagram's content was difficult, leading us to use the Flickr dataset for training, which may not perfectly match the type of content on Instagram, potentially impacting results. To address this, we compiled and manually labeled an Instagram test set. Moreover, the WANN performance is constrained by any limitations in the image and text models, as it needed consistent patterns to emerge from the text and image model to correctly learn how to weigh them.

12 Conclusion

In conclusion, our project navigated through complex data integration and model training challenges to successfully develop a sentiment analysis tool dedicated for Instagram; resulting in a valuable model that provides insights towards more sophisticated social media analytics.

13 Link To Google Colab Notebook

`https://drive.google.com/drive/folders/1R5SH6jMtmcgYnDWNPZ-ldUeEJtBC1yW6?usp=sharing`

REFERENCES

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification, 2020. URL https://aclanthology.org/2020.findings-emnlp.148.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. 2013. URL https://www.ee.columbia.edu/ln/dvmm/vso/download/visual_sentiment_ontology_FINAL.pdf.

Leo Breiman. Random forests. 2001. URL https://doi.org/10.1023/A:1010933404324.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gale Varoquaux. sklearn.feature_extraction.text.tfidfvectorizer, 2024. URL https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html.

Ganesha Chandrasekaran, Naaji Antoanela, Gabor Andrei, Ciobanu Monica, and Jude Hemanth. Visual sentiment analysis using deep learning models with social media data. 2022. URL https://doi.org/10.3390/app12031030.

CrowdFlower. Image sentiment polarity, 2016. URL https://data.world/crowdflower/image-sentiment-polarity.

Pierre Joseph Delorme. Image preprocessing, 2021. URL medium.com/unpackai/image-preprocessing-6654d1bb4daa.

Geeks for Geeks. Rnn for text classifications in nlp, 2024. URL https://www.geeksforgeeks.org/rnn-for-text-classifications-in-nlp/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015. URL https://doi.org/10.48550/arXiv.1512.03385.

Sherif Hussein. Twitter sentiments dataset. 2021. URL https://doi.org/10.17632/z9zw7nt5h2.1.

Instagram. Instagram launches, 2010. URL about.instagram.com/blog/announcements/instagram-launches.

Keras. Densenet, 2024. URL https://keras.io/api/applications/densenet/.

N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. 2018. URL https://doi.org/10.1016/j.knosys.2018.07.041.

MathWorks. vgg19, 2024. URL https://www.mathworks.com/help/deeplearning/ref/vgg19.html.

Szymon Palucha. Understanding openai's clip model, 2024. URL https://medium.com/paluchasz/understanding-openais-clip-model-6b52bade3fa3.

Pinecone. Zero-shot image classification with openai's clip, 2021. URL https://www.pinecone.io/learn/series/image-search/zero-shot-image-classification-clip/.

Alec Radford, Ilya Sutskever, Jong Wook Kim Gretchen Krueger, and Sandhini Agarwal. Clip: Connecting text and images, 2021. URL https://openai.com/research/clip.

TensorFlow. tf.keras.applications.resnet50v2, 2024. URL https://www.tensorflow.org/api_docs/python/tf/keras/applications/ResNet50V2.

Nithyashree V. Image classification using machine learning, 2023. URL https://www.analyticsvidhya.com/blog/2022/01/image-classification-using-machine-learning/.

Huiru Wang, Xiuhong Li, Zenyu Ren, Min Wang, and Chunming Ma. Multimodal sentiment analysis representations learning via contrastive learning with condense attention fusion. 2023. URL https://www.mdpi.com/1424-8220/23/5/2679.

Chen Xi, Guanming Lu, and Jingjie Yan. Multimodal sentiment analysis based on multi-head attention mechanism. 2020. URL https://dl.acm.org/doi/abs/10.1145/3380688.3380693.