Multimedia Tools and Applications (2026) 85:289
https://doi.org/10.1007/s11042-026-21483-4

1248: RECENT ADVANCES IN INTELLIGENT ENVIRONMENTS

Fuzzy logic-driven confidence aggregation for multimodal sentiment classification

Sara Balderas-Díaz[1] · Gabriel Guerrero-Contreras[1] · Andrés Bueno-Crespo[2] · Raquel Martínez-España[3]

Received: 29 November 2024 / Revised: 3 December 2025 / Accepted: 5 March 2026

Abstract
Computational intelligence focuses on intelligent computer systems that mimic human nature and linguistic reasoning. Sentiment analysis is an area of considerable relevance within computational intelligence. Multimodal sentiment analysis is an extension of textual sentiment analysis, where the sentiments of people's opinions are analysed by including multimedia content in addition to textual content. This mode of sentiment analysis faces multiple problems, as the sentiments of text and multimedia content may be contradictory. In addition, another added factor is the imbalance of the data that these problems suffer from in certain topics, which causes a problem when generating intelligent models. In this paper, we design a novel approach for multimodal sentiment analysis, proposing a new way of labelling tweets, not always prioritising polarized classes but using annotator confidence. Moreover, during this design, an information integration and fusion methodology is proposed for the construction of a metamodel that includes fuzzy logic to perform information weighting according to the confidence of the annotator. This proposal has been applied a public unbalanced dataset of tweets with text and images, with a large unbalance towards the negative class label. Applying the proposed fuzzy methodology, we reached a macro-F1 score of 0.493 for the negative class, 0.681 for the neutral class, and 0.832 for the positive class. The model obtains satisfactory performance since the individual image and text sentiment analysis results are worse, especially the negative class, which in initial image classification achieves an F1 score of 0.08.

Keywords Multimodal sentiment analysis · Fuzzy logic · Confidence aggregation · Deep learning

1 Introduction

The explosion of user-generated information on social media from the 2000s onwards has created a significant need for automated tools to process and analyse this information quickly and effectively. The data initially contained text expressing opinions, feelings, and emotions that could provide valuable insights for companies, governments, and institutions. To obtain

Extended author information available on the last page of the article

Published online: 19 March 2026

valid and applicable knowledge from these opinions, an intelligent environment is needed to validate, analyse, and transform such data into interesting and valuable information. This intelligent environment is enabled by computational intelligence. Among its applications, we can highlight natural language processing and image classification [1]. Within this setting, the work presented in this paper advances computational intelligence by combining deep neural architectures with fuzzy-logic-based confidence aggregation and evaluation. At the same time, it strengthens intelligent environments by delivering a sentiment analysis module that is aware of annotator disagreement and class imbalance, and that exposes calibrated, controllable outputs for downstream decision support. The disciplines of natural language processing (NLP) and artificial intelligence (AI) have brought sentiment analysis, a new area of research in the digital world, into the mainstream [2]. Sentiment analysis emerged in response to the need to understand and measure human emotions and opinions in an increasingly digitalised world. Sentiment analysis seeks to automate the identification and extraction of opinions on specific topics. Also referred to as opinion mining, it, along with emotion recognition, is a form of affective analysis. It is commonly used to assess public mood and opinions. It has gained increasing popularity within research communities, academia, public administration, and the service industry. Emotion recognition refers to the process of identifying human emotions. The ability to recognize the emotions of others varies significantly among individuals. The use of technology to aid in emotion recognition is a relatively recent area of study. Affective computing focuses on the automatic detection of an individual's mood or emotional state [3]. In the early days of social media, people primarily expressed their thoughts through text. However, relying solely on textual data can sometimes hinder accurate sentiment prediction. Today, sentiment analysis approaches have evolved to incorporate not only text but also other modalities, such as visual and audio data. As a result, multimodal sentiment analysis (MSA) integrates multiple modalities, extending beyond text- or image-based sentiment analysis. Recent research has focused on recognising sentiment in multimedia content by leveraging multimodal cues, including visual, audio, and textual information. Consequently, the Internet has evolved from a text-based platform into a multimedia-driven one, driving significant advancements in sentiment analysis to support a wide range of applications [4].

On social media, people post text and images to convey their opinions. However, it must be considered that, at times, the text and image may be unrelated or the image may contradict the text, both in terms of information and sentiment [5]. Another aspect to consider is the diverse use of social media. Some social media platforms prioritise image sharing over text, such as Instagram. On these social media platforms, image content tends to be more important than text. In contrast, other social networks, such as X (formerly Twitter), prioritise text over images. In this case, text becomes more relevant than the image, which may not have much in common with the accompanying textual content [6]. Another important factor to consider is that an image's emotional impact may vary depending on the viewer's cultural background. In contrast, text tends to be more direct and is less influenced by the evaluator's perspective [7].

All of the above factors should be considered in MSA approaches to achieve optimal performance. Thus, when designing an MSA approach, machine learning and deep learning are widely used artificial intelligence techniques for building MSA models. These techniques are flexible and enable the adaptation to context and various problem factors to optimise model performance [3, 8].

In this study, we propose a novel approach using an information fusion and integration metamodel to address the MSA problem. Our methodology incorporates fuzzy logic to account for uncertainty arising from the ambiguity in image or text classifications. In the proposed approach, we fuse deep learning and machine learning techniques to extract the most relevant features from text and images for effective sentiment analysis. Additionally, we enhance the information by incorporating text metadata and image colour features to improve sentiment representation and emphasis. The proposed approach is evaluated on a public dataset of tweets containing both images and text. During the development of the methodology, text and image sentiment analysis models are integrated and merged, considering the importance of each element based on the performance of individual models. The key findings of this study are as follows:

- Design of a novel approach to multimodal sentiment analysis.
- Propose a new way of labelling tweets, not always giving priority to polarized classes but using the confidence of the annotators.
- Design of an information integration and fusion methodology for the construction of a metamodel.
- Inclusion of fuzzy logic to weight the confidence of the annotators.

The paper is structured as follows: Section 2 reviews related work on MSA. Section 3 details the methodology, data, and models used in the proposed approach. Section 4 presents the experiments conducted and their analysis. Finally, Section 5 outlines the conclusions and directions for future research.

2 Related work

In recent years, MSA has advanced through the integration of text, images, and other modalities to yield a deeper understanding of sentiment. Approaches incorporating deep learning and attention mechanisms have been developed to address the complexities of sentiment analysis across modalities, such as ambiguity and alignment issues. Techniques for intermediate, late, and hybrid fusion have leveraged BERT and DenseNet201 to combine text and image features, respectively [9]. In intermediate fusion, features from both modalities are integrated early to form a unified vector, while late fusion maintains modality-specific features until predictions are generated and combined through weighted averaging. Hybrid fusion seeks to balance these approaches by allowing for predictions from individual features as well as from a merged feature vector.

Further refinements in fusion techniques have employed cross-attention mechanisms, such as in LXMERT-MMSA, where semantic information alignment between text and images enhances accuracy in detecting sentiment and sarcasm [10]. Additionally, weighted fusion strategies that emphasize embedded textual cues have demonstrated success in visual sentiment analysis (VSA) by incorporating Xception for image features and RoBERTa for text, achieving more nuanced sentiment detection [11]. Incorporating colour cues alongside text and image data has also been shown to improve traditional sentiment classification through image-colour-coding information (ICCI) models, revealing the importance of non-verbal cues in sentiment detection [12].

Ensemble models have shown effectiveness through feature-level fusion and weighted voting strategies. A soft voting ensemble using BiLSTM for textual analysis and EfficientNetB7 for images adjusts the weights assigned to each modality based on classifier performance, adapting dynamically to the specificities of each modality and enhancing robustness in diverse sentiment contexts [13]. Another model, MuAL, combines cross-modal attention with difference loss to further improve modality alignment. The use of difference loss reduces redundancy by introducing orthogonality constraints in text and image embeddings, effectively isolating key sentiment cues from noise [14].

Explorations in cross-instance representation learning, such as the cross-instance graph neural network (CIGNN), highlight the potential of graph-based methods in MSA. CIGNN constructs graphs based on semantic similarity and co-occurrence relationships, bridging gaps across text and image modalities and enhancing the model comprehension of complex inter-instance relations [15]. Frameworks, such as the Context-Sensitive Multi-Tier Deep Learning Framework (CS-MDF), enhance MSA by utilizing CNNs for text, 3D-CNNs for visual data, and openSMILE for audio to extract multimodal features. The multi-tier structure of CS-MDF with BiGRU and GradCAM introduces a context-sensitive approach to sentiment classification, improving interpretability and achieving high performance across various datasets, while underscoring limitations in computational efficiency [16].

Fuzzy logic has been increasingly incorporated to manage ambiguity and enhance model interpretability [17]. Fuzzy Attention Fusion-based MSA (FFMSA) exemplifies this integration by utilizing unsupervised contrastive learning and fuzzy c-means clustering to dynamically manage the interaction of weak and strong sentiment cues across modalities. The reliance on fuzzy attention in FFMSA enables a more adaptable interaction between text, audio, and visual features, which has shown promise in adjusting modality influence based on emotional relevance [18]. Adaptive neuro-fuzzy systems have also been applied to enhance emotion recognition in multimedia contexts, demonstrating that fuzzy logic can efficiently handle uncertainty in sentiment cues while providing high-level interpretability in complex, multimodal datasets [19].

Finally, fuzzy-deep neural networks (Fuzzy-DNN) and the Fuzzy Sentiment Dimension (FSD) have advanced MSA by creating a more refined approach to handling sentiment ambiguity. The Fuzzy-DNN integrates a fuzzy layer with a dual attention mechanism, allowing the model to focus selectively on salient features across modalities, thus refining the balance between redundant and critical features [20]. The FSD model, designed specifically for social network sentiment analysis, introduces a multidimensional approach to sentiment categorization, which transitions smoothly between sentiment degrees from highly negative to positive, moving beyond traditional classifications. The adaptability of this model allows for real-time linguistic adjustment and dynamic membership functions, offering significant depth for analysing sentiments at multiple levels [21].

Despite these advances, existing fuzzy-based MSA approaches mainly apply fuzzy logic at the feature and decision level, for example by fuzzifying attention weights, embeddings, or sentiment dimensions computed from model outputs [17–19]. In contrast, our methodology applies fuzzy logic directly to human annotations and labels. We first derive fuzzy membership degrees from multi-annotator agreement for each modality, then propagate these degrees through confidence-aware sampling and sample-weighted training, and finally incorporate them into a fuzzy evaluation layer with δ-tolerant refinement and WAI metrics that treat near-miss predictions as partially correct. This label- and confidence-centred use

of fuzzy logic is motivated by the inherent ambiguity in human perception of sentiment in social media content, as reflected in the annotator disagreement and class imbalance of the MVSA-Multiple dataset.

3 Methodology for multimodal sentiment analysis

The proposed methodology (Fig. 1), integrates textual and visual modalities to perform sentiment classification of tweets, combining feature extraction and confidence-aware techniques. Initially, image and text annotations are processed separately. For images, preliminary probabilities for positive, neutral, and negative sentiment categories are obtained using the Xception model, complemented by RGB and HSV colour features. For text, contextual embeddings are computed with DeBERTa and simple metadata features (e.g., punctuation counts, length). For each modality, annotator agreement yields a confidence score that is later aggregated to the tweet level. To mitigate class imbalance, we apply ADASYN [22], which synthesizes minority samples by interpolating among nearest neighbours, focusing on sparse or difficult regions rather than duplicating examples. We additionally assign synthetic-sample confidences using a distance-weighted scheme that reflects proximity to real data.

Finally, a deep neural network (DNN) is trained on the rebalanced dataset using confidence scores as sample weights, so higher-certainty instances contribute more to learning. In addition, as shown in Fig. 1, separate weights are applied to image features (W_i) and text features (W_t), enabling the assessment of the relative contribution of text and image features to sentiment classification in the multimodal analysis. Predictions are further refined through a fuzzy evaluation mechanism, which accounts for ambiguities in the annotations by adjusting the final sentiment classification based on confidence-aware rules.

In our methodology, fuzzy logic is introduced to capture the ambiguity and uncertainty of human emotional expression in multimodal social media content, and is specifically tailored to two characteristics of the MVSA-Multiple dataset. First, each tweet has three independent annotators per modality, and disagreement is frequent, especially for negative and neutral labels and when text and image convey divergent cues. If these annotations

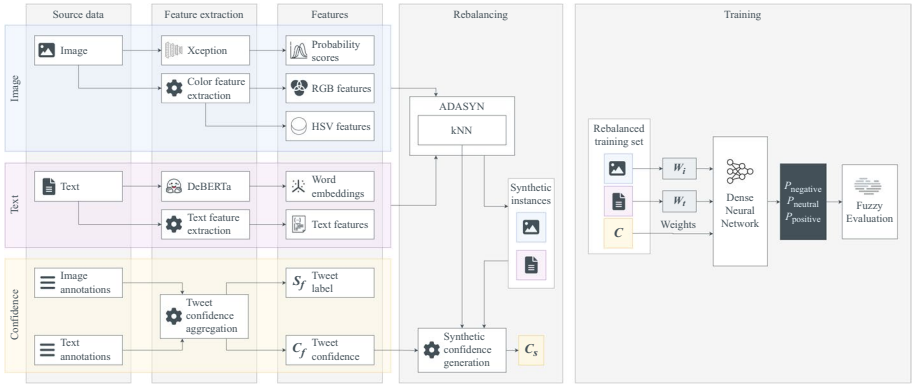


Fig. 1 Proposed methodology for multimodal sentiment classification, showcasing the integration of textual and visual modalities with feature extraction, rebalancing, and confidence-aware techniques

are collapsed into a single crisp label, all information about the degree of agreement and the relative reliability of each modality is lost. Fuzzy membership degrees derived from annotator agreement allow us to represent this uncertainty explicitly and to weight samples and modalities accordingly. Second, the dataset is imbalanced, so minority classes are both scarce and often low-confidence. By combining fuzzy confidences with ADASYN and with a fuzzy evaluation layer, we generate synthetic samples that inherit plausible confidence values, give greater importance to high-confidence minority instances during training, and evaluate near-miss predictions for rare classes through a δ-tolerant fuzzy metric. In this way, fuzzy logic is used to propagate human uncertainty coherently through aggregation, training, and evaluation, rather than merely fuzzifying internal model features.

3.1 Dataset preparation

The MVSA-Multiple dataset (Multi-view Social Data) was used, initially containing 19,600 image-text pairs labelled independently by three annotators [23]. Each pair was annotated separately for sentiment (positive, negative, or neutral), allowing for cases where sentiment in the image differed from that in the text. Only instances where at least two annotators agreed on the sentiment were retained. As a result, pairs with highly mixed annotations (e.g., one positive, one neutral, one negative) or directly contradictory sentiment between text and image (e.g., positive text with a negative image) were excluded from the splits used for training, validation, and testing. This approach prioritizes sentiment consistency across modalities, facilitating a more stable training process [24]. At the same time, we keep pairs where the sentiments expressed by text and image are only mildly divergent (for instance, a positive text with a neutral image). These residual cross-modal discrepancies are not discarded but are later resolved by the confidence-based aggregation scheme described in Section 3.2. After filtering, the dataset comprised 17,027 pairs: 13,620 for training, 1,702 for validation, and 1,705 for testing, structured into an 80-10-10 split.

3.2 Confidence-based sentiment label aggregation

Confidence aggregation drives label assignment and is propagated through fusion and training via fuzzy weighting, allowing the model to account for annotation uncertainty in both the labels and learning.

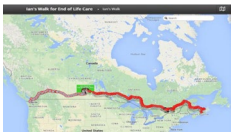
For each sample, whether text or image, the three annotations are analysed to determine a predominant sentiment label and an associated confidence level. When all three annotators agree on the sentiment, as in ID 2502 in Table 1, the sentiment label S is assigned as positive with a confidence level $C = 1.0$, indicating full consensus. In cases where two annotators agree on one sentiment and the third annotator differs, such as ID 2499 in Table 1, the majority sentiment is assigned. In this instance, the final sentiment label S is again positive, but with a reduced confidence level of $C = 0.67$.

To determine the final sentiment label and confidence score for each image-text pair, a fuzzy aggregation approach is implemented. When both modalities indicate the same sentiment, the final confidence score is calculated as a weighted average of their individual confidences. Specifically, the weights w_{t} and w_{i} are computed relative to the sum of both confidences (1).

Table 1 Examples of sentiment annotations in texts with calculated confidence levels

ID	Text	Annotations	Label (S)	Confidence (C)
2502	*"I think it's time for change" - Ana Commit to Vote: #Generation-Trudeau #SFU #LPC #elxn42*	positive positive positive	positive	1.0
2499	*Knocked doors with the venerable #Team-Trudeau #lpc candidate @kylejpeterson this aft in my hometown, Aurora! #elxn42*	positive neutral positive	positive	0.67

Table 2 Example of sentiment agreement across modalities, detailing individual annotations, sentiment labels, and aggregated confidence scores for an image-text pair

ID	Image/Text	Annotations	S_i/S_t	C_i/C_t	S_f	C_f
2620		positive neutral positive	positive	0.67	positive	0.87
	@Justin_Ling I'm walking across #Canada raising awareness for #palliativecare #seniorcare. Need it on #elxn42 agenda	positive positive positive	positive	1.0		

$$w_t = \frac{C_t}{C_t + C_i}, \quad w_i = \frac{C_i}{C_t + C_i} \tag{1}$$

where C_t and C_i represent the confidence scores for the text and image modalities, respectively.

The final confidence score C_f for the combined sentiment label is then derived using (2).

$$C_f = (C_t \times w_t) + (C_i \times w_i) \tag{2}$$

From a fuzzy-set perspective, these confidence scores C_t, C_i and C_f are interpreted as membership degrees in the unit interval for the corresponding sentiment class, derived directly from the discrete agreement counts of the three annotators. With exactly three annotators per instance, full consensus and two-out-of-three agreement already map naturally to membership degrees 1.0 and 2/3, so we do not introduce additional parametric shapes such as triangular, trapezoidal or Gaussian membership functions, which would add free parameters without improving interpretability.

Table 2 provides an example of an image-text pair where both modalities are annotated with a positive sentiment. In this case, the text has a high confidence level ($C_t = 1.0$) due to unanimous agreement among annotators, while the image has a mod-

erate confidence level ($C_{\mathrm{i}} = 0.67$). According to (1), the text receives a higher weight ($w_{\mathrm{t}} = 0.6$) due to its greater confidence level compared to the image, which has a weight of $w_{\mathrm{i}} = 0.4$. Consequently, the overall confidence score for the image-text pair is calculated as $C_{\mathrm{f}} = (1.0 \times 0.6) + (0.67 \times 0.4) = 0.87$, as per (2).

When text and image disagree, the final label S_{f} is taken from the modality with higher annotator-derived confidence. If confidences are equal, polar classes (positive/negative) are preferred over neutral. Therefore, disagreement is resolved by confidence rather than enforcing polarity unconditionally.

Moreover, the final confidence is downweighted by a discount factor d that increases with the difference between text and image confidences (3).

$$d = \frac{1 + |C_{\mathrm{t}} - C_{\mathrm{i}}|}{2} \tag{3}$$

Accordingly, the aggregated confidence C_{f} is given by (4), which incorporates the discount factor d from (3). In this way, the contradictory sentiments between text and image that remain after the dataset filtering step are handled by trusting the more reliable modality, while the discount factor reduces the tweet-level confidence C_{f} whenever the two modalities diverge. As a result, samples with cross-modal inconsistencies are not discarded, but their lower confidence makes them influence training and evaluation proportionally less than high-agreement cases.

For instance, in the case shown in Table 3, the text modality displays a confidence of 1.0 for a positive sentiment, while the image modality shows a confidence of 0.67 for a neutral sentiment. Using (4), the final confidence score is computed as $C_{\mathrm{f}} = ((1.0 \times 0.6) + (0.67 \times 0.4)) \times 0.665 \approx 0.58$.

$$C_{\mathrm{f}} = ((C_{\mathrm{t}} \times w_{\mathrm{t}}) + (C_{\mathrm{i}} \times w_{\mathrm{i}})) \times d \tag{4}$$

Bringing these elements together, confidence aggregation plays a central role in how sentiment labels are defined in this study. First, at the modality level (text and image), the degree of agreement between the three annotators is mapped to a fuzzy confidence value $C \in \{1.0, 0.67\}$ that quantifies how strongly an instance belongs to a sentiment class. Sec-

Table 3 Example of sentiment disagreement between text and image modalities, showcasing the adjustment of final confidence scores and sentiment labels

ID	Image/Text	Annotations	$S_{\mathrm{i}}/S_{\mathrm{t}}$	$C_{\mathrm{i}}/C_{\mathrm{t}}$	S_{f}	C_{f}
2624		positive neutral neutral	neutral	0.67	positive	0.58
	My folks have their @ MattNDP sign up for #Winnipeg South Centre! @ThomasMulcair #elxn42	positive positive positive	positive	1.0		

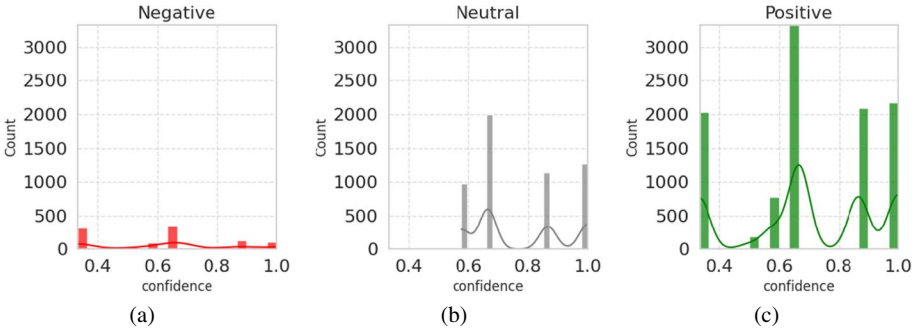


Fig. 2 Frequency distributions of instances based on confidence for negative, neutral, and positive labels

ond, at the tweet level, we aggregate the text and image confidences into a single tweet-level confidence C_f using (1)–(4). This aggregation reinforces the label when both modalities support the same sentiment and downweights cases where the modalities disagree, by combining the higher-confidence modality with a discount factor. The resulting pair (S_f, C_f) is the final fuzzy sentiment label for each tweet and is then propagated through the rest of the pipeline, where S_f is used as the crisp class label and C_f is used as a sample weight during training, as a basis for synthetic confidences, and as the instance weight in the fuzzy evaluation metrics. This approach aligns with state-of-the-art fusion strategies in multimodal sentiment analysis, which leverage adaptive weighting and uncertainty-aware mechanisms to integrate textual and visual information more effectively [8].

Figure 2 shows confidence distributions by label. Negative is less frequent and more dispersed, underscoring the difficulty in achieving annotator consensus for this label; neutral is largely absent at low confidences due to a threshold effect that assigns low-agreement cases to polar classes; positive is most frequent and peaks at high confidence, indicating stronger agreement.

3.3 Textual feature engineering and contextual embeddings

For text sentiment predictions, a hybrid approach was employed [25], combining metadata extraction with contextual text analysis to enhance classification precision. The metadata features considered include text length and word count, which serve as indicators of detailed or emotional expression. Uppercase usage and its ratio within the text are analysed as they often correlate with emphasis or strong emotion. Similarly, the presence of exclamation marks is used to signal enthusiasm or frustration. Additionally, word length ratios are examined to reflect vocabulary complexity and tone, contributing further insights into the emotional and semantic characteristics of the text. Since no outliers were present, MinMax scaling was applied to the metadata to ensure proportional feature contributions in the model.

Following metadata extraction, the text was preprocessed to remove noise and enhance interpretability:

- All text was converted to lowercase to maintain uniformity.
- Numbers and alphanumeric characters, generally lacking inherent sentiment, were removed.

- URLs and user mentions were removed.
- Consecutive repeated letters were trimmed to their base form improving readability.
- Hashtags were expanded to their full meaning. For example, "#sunnymondays" was translated to "sunny Mondays," preserving context and sentiment nuances, as in "#JustinImProudOfYou" to "justin I am proud of you."
- Misspelled words were corrected, and internet slang was translated to formal equivalents (e.g., "tbh" - "To Be Honest," "fomo" - "Fear of Missing Out").
- Punctuation and single-character words, which often add little to sentiment, were removed.

After preprocessing, contextual word embeddings were generated using the DeBERTa model [26], a decoding-enhanced BERT with disentangled attention. The `microsoft/deberta-base` pretrained model and tokenizer were employed. The processed texts were tokenized with padding and truncation applied to a maximum length of 512 tokens. Embeddings were then obtained by passing these tokenized inputs through DeBERTa, extracting the mean of the last hidden state for each sequence to create fixed-size vector representations [27], which were subsequently concatenated with metadata features for a comprehensive input representation.

This hybrid technique leverages metadata to capture indirect emotional cues, while embeddings provide a deeper semantic context, resulting in a holistic view of sentiment.

3.4 Image-based sentiment analysis

3.4.1 Deep feature extraction with Xception

The Xception architecture is based on the Inception network model, emphasizing efficiency and accuracy using separable convolutions in depth [28]. Unlike traditional CNN architectures that stack layers sequentially, Xception employs wider and shallower layers, introducing multiple convolution operations within the same layer, and concatenates their results. This design reduces computational cost while maintaining strong feature extraction capabilities. The architecture works with a 3-channel input format (RGB) with an image size of 299x299.

Xception consists of 36 convolutional layers organized into three main stages: entry flow, middle flow, and exit flow. During the entry flow, the network captures low-level features, while the middle flow, repeated eight times, focuses on deeper, more abstract features. Finally, the exit flow combines and processes this information for prediction. The fully connected layer is replaced with a GlobalAveragePooling2D (GAP) layer, which reduces each feature map to a single value by computing the average of all its activations [29]. GAP layers help reduce overfitting by minimizing the number of trainable parameters while maintaining spatial information from the feature maps (Fig. 3).

The Xception model is initialized with weights pre-trained on ImageNet, providing a foundation of generalized features useful for a wide range of image recognition tasks. To fine-tune the model for custom classification, we freeze the first 20 layers, preserving their pre-trained knowledge, while allowing the remaining layers to adjust during training. On top of the Xception backbone, we add custom classification layers tailored for the specific task, a GAP layer to condense feature maps into meaningful global representations, a dense

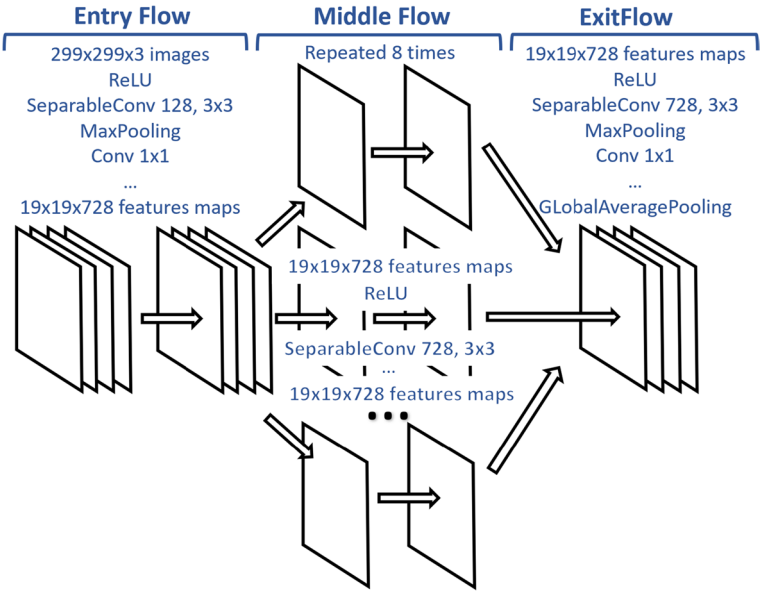


Fig. 3 Architecture of the Xception model

layer with 128 neurons and ReLU activation to learn complex patterns, and a final dense layer with 3 neurons and softmax activation to produce class probabilities for a 3-class classification problem. Between each dense layer, a dropout layer has been included at a rate of 20% to mitigate overfitting.

The model is compiled using an optimised Adam with a learning rate of 1e-4, which balances speed of convergence and stability. A loss function categorical cross-entropy is used, as we are dealing with a multi-class classification problem, and as a metric during training, accuracy is used to evaluate the proportion of correct predictions during training and validation. This configuration ensures an efficient transfer of features from the pre-trained Xception model via ImageNet, and adaptation to our dataset.

3.4.2 Colour feature extraction

Colour theory has been applied in various forms of art and design. Over time, the application of colour theory has become more scientific and data-driven, especially with advances in image processing and sentiment analysis technologies [30]. Psychology and art theory have demonstrated the importance of image colour, which is crucial for expressing feelings [31]. Recent developments have used machine learning models to predict how certain colour combinations influence emotional responses, making this field more accurate and applicable for businesses [32]. Different colours evoke distinct emotional responses; for instance, warm colours like red, orange, and yellow tend to generate positive energy and feelings. Therefore, these basic visual features should be taken into account in MSA [33].

Thus, given the importance of the colour of an image on sentiments, in this study we propose to use the colour information of the images as extra information to improve sentiment classification. Approaching the extraction of colour information from an image can be

done by focusing on two perspectives. On the one hand, extracting the RGB (Red, Green, Blue) colour channels and on the other hand extracting the HSV (Hue, Saturation, Value) values. An initial approach would be to calculate the average of the colours of all channels, for all pixels of the images. However, this faces the problem that all images must have the same dimension in order to obtain equal colour metadata for all images. Another approach would be to focus on the central part of the images and obtain the most frequent colours in that centre. However, this means that if an image does not have the main information in the centre, information would be lost. Therefore, based on these initial ideas, we propose an intermediate approach. The methodology of extracting colour information consists of studying for all RGB and HSV colour channels and analysing in all images which are the most significant colours in these channels. By extracting all channels from all images and analysing the frequencies of values in those channels (RGB and HSV), the most frequent $'x'$ values are obtained, with a significant difference from the rest of the values. Thus, we manage to obtain the $'x'$ most frequent values of all colour channels that can provide relevant information to the sentiment analysis through the images.

In the present study, the values of the most frequent channels with significant differences have been set to 10. Therefore, for each image, 10 values are obtained for each RGB channel and 10 values for each HSV channel. This adds a total of 60 information features to the image sentiment classification.

3.5 Integrated multimodal sentiment metamodel

The proposed multimodal metamodel integrates features extracted from both text and image modalities using a deep neural network (DNN), as depicted in Fig. 4. The architecture consists of an input layer, three fully connected hidden layers with dropout, and an output layer designed for sentiment classification.

The input layer processes a concatenated feature vector composed of weighted text and image representations. The text feature vector X_t includes both contextual embeddings obtained from DeBERTa and text metadata (e.g., punctuation usage, text length, capitalisa-

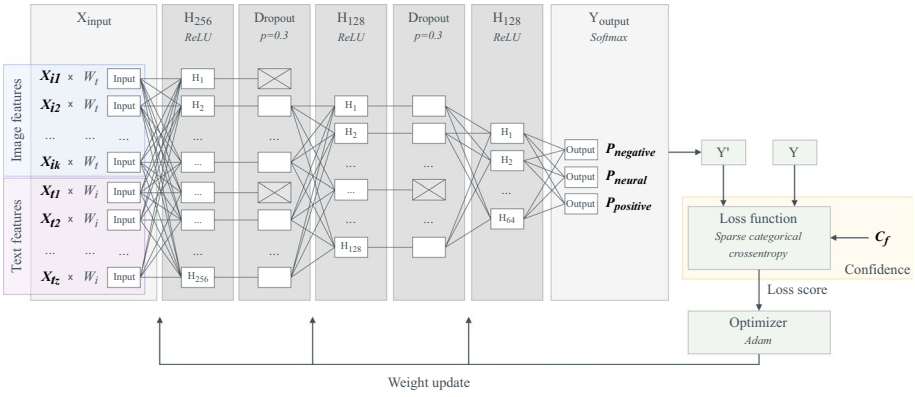


Fig. 4 Architecture of the proposed deep neural network. The model integrates weighted textual and visual features, applying confidence-aware adjustments to enhance classification accuracy. The network consists of three fully connected layers with ReLU activation, dropout regularization to prevent overfitting, and a final softmax output layer for sentiment classification. Confidence scores (C_f) are incorporated as sample weights to prioritize high-certainty instances

tion), and is scaled by a modality-specific weight W_t. Similarly, the image feature vector X_i, which consists of deep features extracted via the Xception model along with colour statistics in RGB and HSV spaces, is scaled by its corresponding weight W_i. These weights, set to 1.5 for text and 0.9 for images, were determined through a grid search on the validation set to optimise the macro-F1 score, exploring ranges of $W_t \in \{0.5, 1.0, 1.5, \ldots 3.0\}$ and $W_i \in \{0.1, 0.2, 0.3, \ldots, 1.0\}$. The final values reflect the greater contribution of textual content in multimodal sentiment classification in social networks like X [34], where text often conveys more explicit sentiment cues and images complement this information. In this setup, the multimodal classifier receives as input a single feature vector obtained by concatenating the weighted text representation X_{t} and the weighted image representation X_{i}, while the tweet-level confidence C_{f} is used only as a sample weight in the loss function and is not concatenated as an input feature.

The first hidden layer contains 256 neurons and uses the ReLU activation function. To mitigate overfitting, it is followed by a dropout layer with a rate of $p = 0.3$, randomly deactivating 30% of neurons during training. The second hidden layer consists of 128 neurons, again with ReLU activation and a similar dropout configuration. The third hidden layer includes 64 neurons, maintaining the ReLU activation function for consistent feature extraction.

The output layer is a fully connected dense layer with three neurons corresponding to the sentiment classes: negative, neutral, and positive. A softmax activation function is applied, producing a probabilistic distribution over the sentiment classes. These probabilities, which represent the confidence of the model for each class, are compared against the ground truth labels to compute the loss using sparse categorical cross-entropy. Simultaneously, confidence scores are incorporated as sample weights, reflecting the certainty of each instance and allowing the model to prioritize higher-confidence samples during training. The model optimisation is performed using the Adam optimiser, which adjusts weights iteratively to minimize the loss function.

3.6 Balancing sentiment classes with confidence-aware sampling

The final dataset exhibits a noticeable imbalance across sentiment labels. Positive annotations dominate both image and text modalities, followed by neutral and negative annotations, which are significantly less frequent. This trend persists after aggregating the sentiment labels at the tweet level (see Subsection 3.2), as shown in Table 4.

The ADASYN method is employed to address this imbalance by generating synthetic samples for the minority sentiment classes [22]. ADASYN uses the k-nearest neighbours (kNN) algorithm to identify the local distribution of each minority instance in the feature space. For each instance, synthetic samples are generated by interpolating between the instance and one of its k-nearest neighbours.

Table 4 Sentiment distribution across image, text and tweet aggregation annotations

Sentiment	Image	Text	Tweet aggregation
Positive	9,365	9,540	10,583
Neutral	6,813	6,431	5,382
Negative	849	1,056	1,062

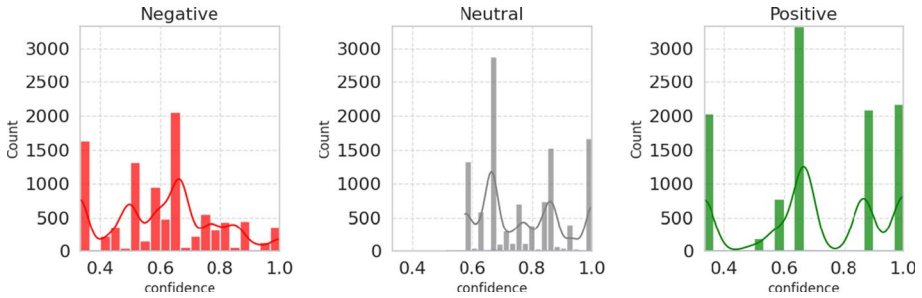


Fig. 5 Confidence distributions for sentiment labels after applying ADASYN to balance the dataset, showing the effect of synthetic samples

In this study, it is proposed to enhance this approach by including the computation of confidence values for the synthetic samples. Rather than assigning a uniform or arbitrary confidence score, the confidence of a synthetic sample is calculated as the weighted average of the confidence scores of its k-nearest neighbours. Formally, let x_i represent a minority instance, and let $N_k(x_i)$ denote the set of its k-nearest neighbours. The confidence score $C_{\text{synthetic}}$ for a generated synthetic sample is defined in (5).

$$C_{\text{synthetic}} = \frac{\sum_{x_j \in N_k(x_i)} w_j \cdot C_j}{\sum_{x_j \in N_k(x_i)} w_j} \tag{5}$$

where C_j is the confidence score of the neighbor x_j, and w_j is a weight inversely proportional to the distance between x_i and x_j.

The hyperparameters for ADASYN were selected through a grid search on the validation dataset, with the goal of optimizing the macro-F1 score. Specifically, the number of neighbours (k) was set to $k = 5$ for text data, $k = 7$ for image data, and $k = 3$ for the multimodal dataset. The variation in k reflects the differing characteristics of the modalities.

This weighted averaging preserves confidence distributions of the original data, as shown in Fig. 5, thereby minimizing the risk of bias in the training process.

3.7 Fuzzy refinement of multimodal sentiment predictions

The fuzzy evaluation process combines the outputs of the DNN with the pre-calculated confidence scores from the tweet to refine sentiment predictions applying the flexibility that fuzzy logic allows us. This approach adapts dynamically to agreement and discrepancies between the model predictions and the confidence-based labels.

The inputs to this process include the predicted probabilities of the DNN for each sentiment class ($P_{\text{negative}}, P_{\text{neutral}}, P_{\text{positive}}$), and the aggregated confidence score for the tweet (C_f).

When the predicted class of the DNN, identified as the one with the highest probability ($P_{\text{predicted}}$), matches the agreed sentiment label (S_f) the prediction is considered correct. The combined confidence score (C_f) is directly assigned as the weight for the evaluation.

If the predicted class does not align with the agreed sentiment label, the discrepancy is systematically assessed by calculating the difference between the highest predicted prob-

ability ($P_{\text{predicted}}$) and the probability corresponding to S_f (P_{S_f}). When this difference is less than a predetermined threshold, it indicates a lack of decisive confidence in the predicted class. The threshold δ enables a flexible decision mechanism that allows the model prediction to be adjusted when the confidence gap between the predicted class and the agreed sentiment label S_f is small. In probabilistic terms, δ defines the maximum allowed gap between the model-assigned probability of the predicted class and the probability assigned to the confidence-based label S_{f} that we still treat as a near miss, so larger values of δ therefore tolerate greater discrepancies before a prediction is considered fully incorrect. Specifically, if the difference $P_{\text{predicted}} - P_{S_f}$ is less than δ, the model prediction is overwritten with S_f, and the sample weight is adjusted as $W_{\text{adjusted}} = C_f \times (1 - (P_{\text{predicted}} - P_{S_f}))$. Otherwise, when the probability gap is greater than or equal to δ, the model prediction is kept unchanged and the instance is evaluated with its original weight C_{f}, so the fuzzy evaluation coincides with the crisp one for that sample. In practice, this creates two regimes for discrepant cases. For strong mismatches, the model prediction is fully preserved. For near-miss predictions, the final sentiment is reset to the confidence-based label S_{f} and the reduced weight $W_{\text{adjusted}} < C_{\text{f}}$ assigns partial credit. This formulation captures and modulates the uncertainty arising from ambiguous predictions, providing a smooth correction mechanism in borderline cases.

From a fuzzy-set perspective, this procedure can be viewed as a small fuzzy rule base defined on the discrete universe of tweets. Each instance *k* has a confidence-based label $S_{\text{f}}(k)$ and an associated membership degree $W_k \in (0, 1]$, which is C_{f} in the crisp regime and W_{adjusted} in the near-miss regime. For a given class *i*, the fuzzy set of ground-truth instances is therefore characterised by the membership function $\mu_i^{\text{true}}(k) = W_k$ if $S_{\text{f}}(k) = i$ and $\mu_i^{\text{true}}(k) = 0$ otherwise. In (13)-(16) the weighted counts $TP_{i,\text{weighted}}$, $FP_{i,\text{weighted}}$ and $FN_{i,\text{weighted}}$ are thus sums of membership degrees rather than raw counts. In the WAI in (17), the term $\min(P_k^{\text{pred}}, P_k^{\text{true}})$ plays the role of a fuzzy conjunction between predicted and target membership, using the minimum operator as a standard *t*-norm.

To illustrate the fuzzy evaluation process, consider a sample with predicted probabilities for the DNN as $P_{\text{negative}} = 0.25$, $P_{\text{neutral}} = 0.35$, and $P_{\text{positive}} = 0.40$. The aggregated confidence score for the tweet is $C_f = 0.85$, and the agreed sentiment label is $S_f = \text{neutral}$.

The predicted class, identified as the one with the highest probability, is $P_{\text{predicted}} = P_{\text{positive}} = 0.40$. This does not align with the agreed sentiment label, $S_f = \text{neutral}$.

To assess the discrepancy, the difference between the highest and second-highest probabilities is calculated as $P_{\text{predicted}} - P_{S_f} = 0.05$. Since this difference is less than a threshold, e.g., $\delta = 0.2$, the prediction is overwritten with the confidence-based label $S_{\text{f}} = \text{neutral}$. The weight is adjusted to account for this uncertainty as $W_{\text{adjusted}} = 0.85 \times (1 - 0.05) = 0.85 \times 0.95 = 0.8075$.

3.8 Performance metrics

To evaluate the performance of the model, a set of complementary metrics is employed: F1 score, Balanced accuracy, the Matthews Correlation Coefficient (MCC), and Cohen's Kappa.

The F1 score, defined in (6), represents the harmonic mean of precision (7) and recall (8) for each class i.

$$F_{1,i} = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}, \tag{6}$$

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \tag{7}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}. \tag{8}$$

where TP_i, FP_i, and FN_i refer to the true positives, false positives, and false negatives for class i, respectively. For multiclass classification problems, the macro-averaged F1 score is used to aggregate the F1 scores of all classes equally, as shown in (9).

$$\text{macro-F1} = \frac{1}{C} \sum_{i=1}^{C} F_{1,i}, \tag{9}$$

where C represents the total number of classes, and $F_{1,i}$ denotes the F1 score of class i. This macro-averaging ensures that performance on minority classes is not overshadowed by the dominance of majority classes.

Balanced accuracy measures the average recall obtained on each class (10).

$$\text{Balanced accuracy} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FN_i}, \tag{10}$$

MCC, defined in (11), is a metric that considers all elements of the confusion matrix (true positives, true negatives, false positives, and false negatives).

$$\text{MCC} = \frac{\sum_{i,j,k} \left(TP_i \cdot TN_k - FP_i \cdot FN_j\right)}{\sqrt{\left(\sum_i P_i \cdot N_i\right) \left(\sum_j P_j \cdot N_j\right)}}, \tag{11}$$

where P_i and N_i denote the total predicted positives and negatives for class i, respectively, and the indices j and k iterate over all classes. MCC ranges from -1 to 1, where 1 indicates perfect correlation between predictions and true labels, 0 represents no correlation, and -1 signifies total disagreement.

Finally, Cohen's Kappa measures the agreement between the predicted and true labels while accounting for the agreement expected by chance (12).

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{12}$$

where p_o is the observed agreement, and p_e is the expected agreement based on chance. Cohen's Kappa ranges from -1 to 1, with values closer to 1 indicating strong agreement

beyond chance, 0 implying agreement equivalent to random guessing, and negative values suggesting systematic disagreement.

To further enhance the evaluation process, fuzzy-oriented metrics are introduced to account for partial correctness and confidence-based adjustments in predictions.

True positives (TP_i), false positives (FP_i), and false negatives (FN_i) for each class i are adjusted using the weights as described in (13).

$$TP_{i,\text{weighted}} = \sum_{k \in TP_i} W_k, \quad FP_{i,\text{weighted}} = \sum_{k \in FP_i} W_k, \quad FN_{i,\text{weighted}} = \sum_{k \in FN_i} W_k \tag{13}$$

These weighted terms are then substituted into the standard definitions of precision (14), recall (15), and F1 score (16).

$$\text{Weighted Precision}_i = \frac{TP_{i,\text{weighted}}}{TP_{i,\text{weighted}} + FP_{i,\text{weighted}}} \tag{14}$$

$$\text{Weighted Recall}_i = \frac{TP_{i,\text{weighted}}}{TP_{i,\text{weighted}} + FN_{i,\text{weighted}}} \tag{15}$$

$$F_{1,i}^{\text{Weighted}} = 2 \cdot \frac{\text{Weighted Precision}_i \cdot \text{Weighted Recall}_i}{\text{Weighted Precision}_i + \text{Weighted Recall}_i} \tag{16}$$

Additionally, a fuzzy agreement metric, the Weighted Agreement Index (WAI) (17), is introduced to measure the alignment between predicted and true membership distributions.

$$\text{Weighted Agreement Index} = \frac{\sum_k W_k \cdot \min(P_k^{\text{pred}}, P_k^{\text{true}})}{\sum_k W_k \cdot P_k^{\text{true}}} \tag{17}$$

Here, W_k is the weight for instance k, P_k^{pred} is the predicted probability for the true class, and P_k^{true} is the true probability (usually 1 if fully correct, adjusted if partially correct).

4 Evaluation and performance analysis

In this section, the results from a crisp and fuzzy point of view are analysed in detail, taking into account the individual text and image models and the multimodal models.

4.1 Crisp evaluation results

4.1.1 Text model performance

The baseline model, relying solely on DeBERTa embeddings, demonstrates moderate performance, achieving a macro-F1 score of 0.476, MCC of 0.301, balanced accuracy of 46.4%, and Cohen's Kappa of 0.295 (Table 5). Among the individual classes (Table 6), the model performs well in identifying the positive class, with an F1 score of 0.73, but struggles significantly with the minority negative class, achieving an F1 score of only 0.16. The neu-

Table 5 Performance comparison across different approaches for text sentiment classification

Approach	macro-F1	MCC	Balanced Acc.	Kappa
Text Only (DeBERTa)	0.476	0.301	46.4%	0.295
Metadata	0.485	0.311	47.1%	0.305
CW loss	0.499	0.316	48.0%	0.311
ADASYN Rebalancing	**0.566**	**0.339**	**56.7%**	**0.339**

Table 6 Class-specific F1 scores across different approaches for text sentiment classification

Approach	F1 (neg.)	F1 (neu.)	F1 (pos.)
Text Only (DeBERTa)	0.160	0.540	0.730
Metadata	0.170	0.550	**0.740**
CW loss	0.210	0.560	0.730
ADASYN Rebalancing	**0.400**	**0.580**	0.720

tral class exhibits moderate performance, with an F1 score of 0.54. This baseline establishes a reference point, highlighting the inherent challenges of class imbalance and the limitations of relying exclusively on embeddings for sentiment classification.

Incorporating metadata features such as exclamation counts and uppercase-to-lowercase ratios slightly improves the results. The macro-F1 score increases to 0.485, MCC rises to 0.311, and balanced accuracy reaches 47.1% (Table 5). Cohen's Kappa also improves slightly, reaching 0.305, indicating a marginal increase in agreement between the model predictions and the ground truth. Class-specific improvements (Table 6) are observed uniformly across all sentiment classes. The F1 score for the negative class increases from 0.16 to 0.17, while the neutral class rises from 0.54 to 0.55, and the positive class improves slightly from 0.73 to 0.74. These consistent gains indicate that the inclusion of metadata features contributes to a general refinement in predictions.

Integrating fuzzy logic into the training process to assign confidence-based weights to the samples results in further overall improvement. This approach achieves a macro-F1 score of 0.499, MCC of 0.316, and balanced accuracy of 48.0% (Table 5). Class-specific F1 scores exhibit varying levels of improvement (Table 6). The negative class shows the most significant gain, increasing from 0.17 to 0.21, while the neutral class improves slightly from 0.55 to 0.56. In contrast, the positive class experiences a slight reduction in performance, with its F1 score decreasing marginally from 0.74 to 0.73. These improvements demonstrate the value of incorporating annotator-derived confidence scores from annotators, which help to minimize the impact of noise and ambiguity in the dataset.

The application of the ADASYN rebalancing technique yields significant gains, particularly in addressing the challenges of class imbalance. As shown in Table 5, this approach achieves a macro-F1 score of 0.566, MCC of 0.339, and balanced accuracy of 56.7%. Additionally, Cohen's Kappa improves to 0.339. The impact of ADASYN is most evident in the negative class (Table 6), where the F1 score increases dramatically from 0.16 in the baseline to 0.40. The neutral class also benefits from ADASYN, with its F1 score improving to 0.58, while the positive class remains stable, achieving an F1 score of 0.72.

4.1.2 Image model performance

Analysing, in contrast, the results of the image sentiment analysis models yield discrete values of 0.366, 0.122, 38.03%, and 0.107 for the metrics F1 score, MCC, balanced accu-

Table 7 Performance comparison across different approaches for image sentiment classification

Approach	macro-F1	MCC	Balanced Acc.	Kappa
Image Only	0.366	0.123	38.03%	0.107
Colour	0.411	0.163	40.69%	0.159
CW loss	0.416	0.167	41.04%	0.163
ADASYN Rebalancing	**0.421**	**0.170**	**43.47%**	**0.169**

Table 8 Class-specific F1 scores across different approaches for image sentiment classification

Approach	F1 (neg.)	F1 (neu.)	F1 (pos.)
Image Only	0.087	0.309	**0.702**
Colour	0.108	0.448	0.677
CW loss	0.121	0.446	0.680
ADASYN Rebalancing	**0.161**	**0.463**	0.640

racy, and Cohen's kappa respectively as can be seen in Table 7. Examining the individual results by class (Table 8) the positive class is acceptably identified with a 0.702 value of F1 score. The neutral class is moderately identified with an F1 score value of 0.309. Finally, the negative class (minority class) is the worst identified with an F1 score value of 0.087. This situation indicates the imbalance problem already known in the text, it still appears with the classification of the images.

After incorporating information on Colour, the results indicated above improve slightly, especially in the minority classes, which was the worst classified by the model (Tables 7 and 8). Thus, the F1 score reaches 0.411 value, the MCC obtains a 0.163, the balanced accuracy reaches a 40.69% and the Cohen's kappa obtains a 0.159 (Table 7). Regarding the individual results, Class-specific F1 scores values, shown in Table 8 improve for the negative and neutral classes, but worsen slightly for the positive class. This is not a negative result, but rather highlights the overfitting suffered in this positive class, to the detriment of the other two. By adding colour as metadata, the model manages to reduce this overfitting in the positive class and obtain a better result in the negative and neutral classes. Actually, the neutral class is the one that benefits the most, since the improvement of the negative class is less progressive.

In the following model, integrating fuzzy logic into the training process to assign confidence-based weights to samples results in a slight improvement overall. Table 7 shows the slight improvement of the values, where the F1 score value is 0.416, the MCC value is 0.167, the balanced accuracy value is 41.04% and the Kappa value is 0.163. Analysing the values in particular (see Table 8), it can be seen that there are also slight improvements in the individual classes, being the negative class the one that improves the most, reaching a value of 0.121 F1 score. The positive class obtained a value of 0.68 and the neutral class a value of 0.446 F1 score.

The application of the ADASYN rebalancing technique produces overall benefits. Thus, Table 7 shows how the F1 score value amounts to a value of 0.421, the MCC value is 0.17, the balanced accuracy value 43.47% and the Cohen's kappa value amounts to 0.169. The individual results shown in Table 8, after the application of ADASYN, show a rebalancing of the results, increasing the improvement in the minority classes (neutral and negative). Thus, the negative classes achieve an F1 score value of 0.161, doubling the improvement over the initial model. Conversely, the neutral class obtains an F1 score value of 0.463 and the majority class (positive class) lowers its value to 0.640.

4.1.3 Multimodal model performance

The application of confidence-weighted (CW) loss and the ADASYN rebalancing techniques has led to improvements in the results of individual sentiment classification models in both text and image. In addition, they also perform a rebalancing of the results, improving classification in minority classes. Now, these two techniques are applied to the multimodal model, which merges all the information provided by text and image. Initially the results of the multimodal model without applying confidence-weighted loss and the ADASYN rebalancing techniques reach an overall performance of 0.535 of F1 score, 0.298 of MCC, 51.6% of balanced accuracy and 0.289 of Cohen's kappa (see Table 9). Analysing this approach individually in the sentiment classification, the F1 score results show a value of 0.382 for the negative class, 0.476 for the neutral class and 0.747 for the positive class. These results substantially improve the individual results of the text and image classification models in the mainly negative class (see Table 10).

By applying the confidence-weighted loss technique, remarkable benefits are obtained. Table 9 shows how the F1 score value reaches a value of 0.553, the MCC value is 0.309, the balanced accuracy value is 53.0% and Cohen's kappa value is 0.308. In particular, the F1 score value for the positive class is 0.728, somewhat lower than the model without the application of the confidence-weighted loss technique. However, the negative and neutral classes, minority classes, obtain an F1 score value of 0.397 and 0.533 respectively, a substantial increase in the neutral class (Table 10).

Taking the ADASYN Rebalancing technique approach, the results are visibly improved. Thus, the F1 score value becomes 0.574, the MCC value is 0.330, the balanced accuracy value is 59.4% and Cohen's Kappa value is 0.329 (Table 9). At the individual class level, the positive class slightly increases its F1 score value to 0.734. On the other hand, the negative and neutral classes increase their F1 score value to 0.464 and 0.524 respectively (Table 10), being remarkable the rise of the negative class (minority class). The negative class obtained an F1 score value of 0.40 for the text-only model and 0.16 for the image-only model. When merging the information this class increases to 0.464 for F1 score value. To assess the robustness of this configuration, we repeated the training of the final multimodal model with ADASYN using three different random seeds. Across these runs, the test macro-F1 was 0.573 ± 0.002 and the balanced accuracy was $59.2\% \pm 0.4\%$, which shows that performance varies only marginally and that the stochastic components of ADASYN and multimodal fusion do not introduce substantial instability. Analysing the training, validation, and test phases, we observe that the model achieves a macro-F1 score of 0.679 during

Table 9 Performance comparison across different approaches for multimodal sentiment classification

Approach	macro-F1	MCC	Balanced Acc.	Kappa
Text + Image	0.535	0.298	51.6%	0.289
CW loss	0.553	0.309	53.0%	0.308
ADASYN Rebalancing	**0.574**	**0.330**	**59.4%**	**0.329**

Table 10 Class-specific F1 scores across different approaches for text and image sentiment classification

Approach	F1 (neg.)	F1 (neu.)	F1 (pos.)
Text + Image	0.382	0.476	**0.747**
CW loss	0.397	**0.533**	0.728
ADASYN Rebalancing	**0.464**	0.524	0.734

training, which decreases to 0.618 in validation and further to 0.574 in the final test set. The moderate drop of approximately 6.1% from training to validation indicates that the model does not excessively memorize training data, while the validation-to-test difference of approximately 4.4% aligns with expected generalisation behaviour. In summary, the results obtained from a crisp point of view, are generally better using the combination and fusion of information as opposed to individual models. Above all, the improvement is focused on the minority class (negative sentiment).

4.2 Fuzzy evaluation results

The fuzzy evaluation introduces flexibility in assessing model performance by allowing partial correctness. This is achieved by defining a threshold, δ, which quantifies the allowable difference between the predicted class probability and the probability of the correct class, and determines when an instance is considered a fuzzy match, thus enhancing interpretability and accounting for ambiguity in the predictions. In all analyses below, we treat δ as an evaluation-time tolerance and perform a validation sweep $\delta \in 0.0, 0.1, 0.2, 0.3, 0.4, 0.5$ with the trained models kept fixed (no retraining across δ); $\delta = 0$ corresponds to crisp evaluation. We select δ^* on the validation split by maximising weighted-F1 subject to preserving agreement at a level $\gamma \in [0.9, 0.95]$, as shown in (18). This procedure links the choice of δ to annotation uncertainty through the WAI, which measures agreement between model probabilities and the confidence-based labels derived from annotators. By requiring $WAI(\delta)$ to remain within 90–95% of the crisp baseline *WAI*(0), we ensure that the fuzzy tolerance does not move the evaluation far from what annotators consider correct, while still allowing partial credit for near-miss cases.

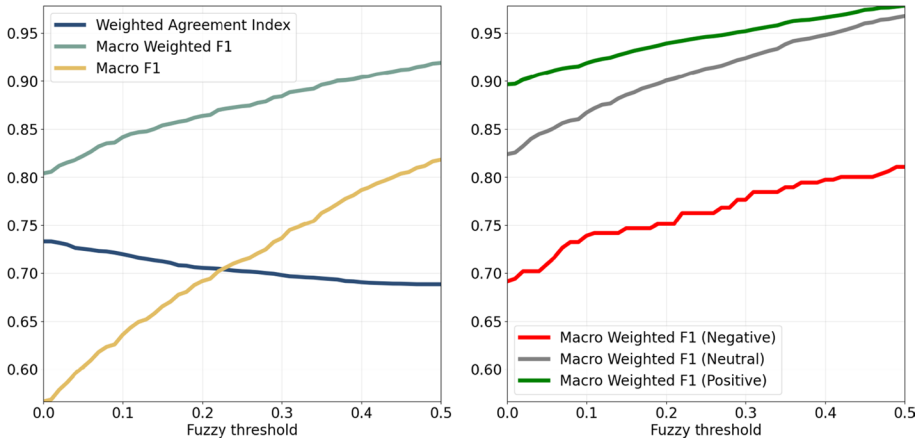
$$\delta^* = \arg \max_{\delta \in \{0.0, 0.1, \ldots, 0.5\}} \mathrm{F1}_{\mathrm{weighted}}(\delta) \quad \text{s.t.} \quad \mathrm{WAI}(\delta) \geq \gamma \cdot \mathrm{WAI}(0) \tag{18}$$

4.2.1 Text-based analysis

Figure 6 illustrates the impact of threshold δ in the model performance for textual sentiment classification. At $\delta = 0$, the macro-F1 score is 0.804, driven by strong performance in the positive class (0.897) and the neutral class (0.824), while the negative class achieves a lower score of 0.691. The WAI at this threshold is relatively high, at 0.733, indicating strong agreement for high-confidence predictions. This baseline reflects the model's strength in identifying dominant sentiments but highlights the challenges associated with the negative class.

As δ increases, the evaluation incorporates partial correctness, leading to consistent improvements in the weighted-F1 score, which reaches 0.919 at $\delta = 0.5$. The WAI decreases marginally to 0.688 at $\delta = 0.5$, reflecting the inclusion of lower-confidence predictions without significant loss of alignment between predicted and true values.

Class-specific weighted-F1 scores show significant gains as δ increases. The negative class improves steadily, reaching 0.811 at $\delta = 0.5$, highlighting the framework's ability to address under-represented classes. The neutral class benefits from fuzzy evaluation as well, rising from 0.824 at $\delta = 0$ to 0.968 at $\delta = 0.5$. The positive class, which starts strong, achieves near-perfect performance with a score of 0.979 at $\delta = 0.5$. The selected operating point is $\delta^* = 0.3$, which lies on a broad plateau of the sensitivity curve and offers a balanced trade-off across classes (weighted-F1 $= 0.884$; negative $= 0.776$, neutral $= 0.924$, positive $= 0.952$).



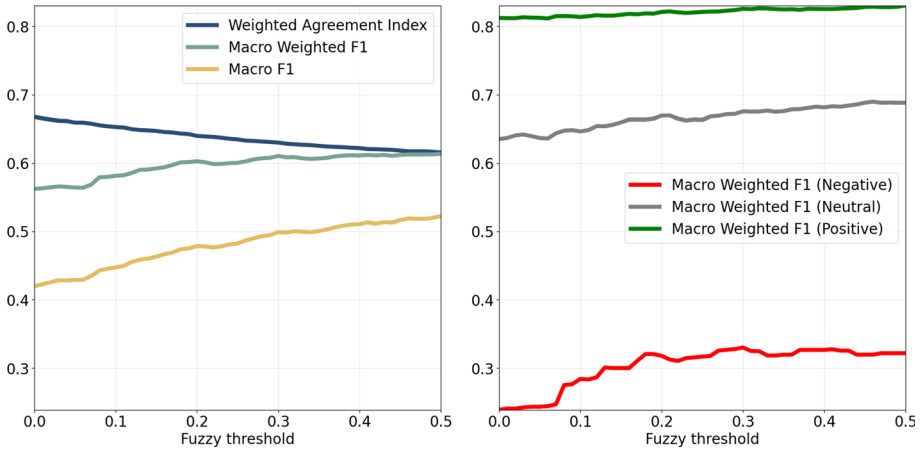
(a) (b)

Fig. 6 Sensitivity to the fuzzy threshold δ for the textual branch

4.2.2 Image-based analysis

Figure 7 summarises the key metrics for the fuzzy evaluation for image sentiment analysis. At $\delta = 0$, the WAI is relatively high at 0.668, indicating a good alignment between predicted and true probabilities for confident samples. The overall weighted-F1 score is 0.562, driven largely by strong performance in the positive class (0.812), while the neutral and negative classes lag behind with scores of 0.635 and 0.239, respectively. As in the text branch, we report a sensitivity curve with $\delta \in \{0.0, \ldots, 0.5\}$ and fixed model parameters.

As δ increases, the evaluation becomes more flexible. This adjustment improves the weighted-F1 score, which peaks at 0.613 around $\delta = 0.5$. The WAI, however, decreases slightly to 0.616, reflecting the inclusion of lower-confidence instances as correct predictions.



(a) (b)

Fig. 7 Sensitivity to the fuzzy threshold δ for the image branch

Class-specific improvements are most notable in the negative and neutral classes. The weighted-F1 score for the negative class steadily rises, reaching 0.322 at $\delta = 0.5$. The neutral class also benefits, with its score increasing from 0.635 at $\delta = 0$ to 0.688 at $\delta = 0.5$. The positive class remains stable throughout, with only marginal improvements, reflecting its dominance in the dataset and the model's robust performance in this class. Consistent with the textual branch, the operating point $\delta^* = 0.3$ provides a stable balance (weighted-F1 $= 0.611$; negative $= 0.327$, neutral $= 0.676$, positive $= 0.825$).

4.2.3 Multimodal analysis

Figure 8 summarises the key metrics for fuzzy evaluation applied to multimodal sentiment analysis. At $\delta = 0$, the WAI achieves a high value of 0.910, reflecting strong alignment between predicted and true probabilities for high-confidence samples. The macro-F1 score at this threshold is 0.639, driven predominantly by robust performance in the positive class with a score of 0.732, while the neutral and negative classes achieve 0.513 and 0.452, respectively. We again report a sensitivity curve over $\delta \in \{0.0, \ldots, 0.5\}$ with fixed model parameters.

As δ increases, it results in consistent improvements in the weighted-F1 score, which peaks at 0.862 at $\delta = 0.5$. The WAI gradually declines as lower-confidence matches are included, stabilizing at 0.825 by $\delta = 0.5$, reflecting a controlled trade-off between inclusivity and precision.

The class-specific weighted-F1 scores exhibit distinct trends as δ varies. The negative class shows steady improvement, reaching 0.518 at $\delta = 0.5$. The neutral class also benefits significantly, rising from 0.513 at $\delta = 0$ to 0.810 at $\delta = 0.5$, showcasing better handling of ambiguous instances. The positive class, consistently strong, improves to 0.906 at $\delta = 0.5$, reflecting its continued dominance across thresholds. The operating point $\delta^* = 0.3$ achieves a balanced trade-off (weighted-F1 $= 0.769$; negative $= 0.493$, neutral $= 0.681$, positive $= 0.832$) and lies within a flat region of the sensitivity curve, which reduces the risk of overtuning to a single threshold. The selection follows the rule weighted-F1 subject to a WAI constraint, avoiding retraining across δ and ensuring comparability.

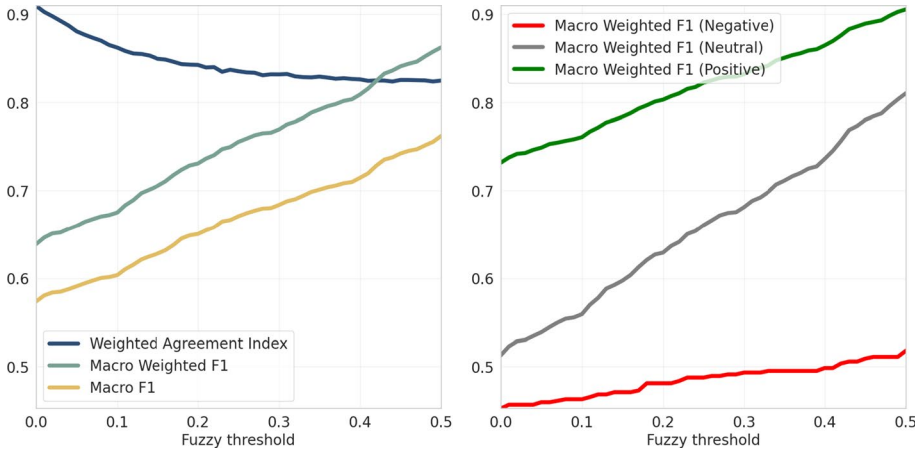


Fig. 8 Sensitivity to the fuzzy threshold δ for the multimodal model

On new datasets, we recommend a validation sweep over $\delta \in [0.0, 0.5]$ and selecting δ^* with the same F1–WAI rule. When validation is limited, robust defaults in the plateau region $\delta \in [0.2, 0.4]$ preserve agreement while capturing near-miss cases. If annotator agreement/confidence differs markedly from our data, reselect δ accordingly and report both views (crisp δ=0 and fuzzy at δ^*).

The examples in Table 11 illustrate how the fuzzy evaluation framework addresses sentiment ambiguity in multimodal scenarios. These instances represent typical challenges where annotators assign diverging sentiment labels to the same input due to contextual subtleties or subjective interpretation of visual and textual cues. By examining the model's output under a fuzzy agreement threshold ($\delta = 0.3$), we gain qualitative insight into its ability to handle uncertainty in sentiment assignment, highlighting both its robustness in cases with a dominant modality and its limitations in highly ambiguous or conflicting situations.

For ID 12774, the image shows a map that offers little emotional or contextual information relevant to sentiment classification. The image annotations reflect this ambiguity, with a mix of neutral and negative sentiments. In contrast, the textual content clearly conveys a negative sentiment by describing a robbery incident. The resulting sentiment label is negative, with high confidence ($C_f = 0.87$). This case exemplifies the greater capacity of textual content to convey sentiment with clarity, especially in scenarios where visual elements lack explicit emotional cues. By prioritising the stronger signal from the text, the framework effectively compensates for the ambiguity present in the image modality.

ID 13974 presents a scenario of ambiguity both within and across modalities. The image, displaying the word "SALE" in bold red text, might convey excitement, but the annotators unanimously assign a neutral sentiment. Meanwhile, the textual content describes an art sale, with annotations split between neutral and positive sentiments. The framework reconciles these differences by assigning a neutral sentiment, which aligns with the stronger agreement in the image modality. However, the confidence ($C_f = 0.58$) is lower due to the disagreement. This case illustrates the complexity of integrating visual and textual cues when the interpretations are may vary among annotators.

Table 11 Examples of instances where fuzzy evaluation successfully identified the final sentiment label with a $\delta = 0.3$

ID	12774	13974	13698
Image			
Annotations	neu., neg., neg.	neu., neu., neu.	pos., neu., pos.
Text	Woman robbed while walking to work in Riverdale #HamOnt #sc	Today's the day!! One day #artsale at my studio. Drop by 11-4pm at 1395B Welly. Select pieces only. #WellingtonWest	@RobbyivyIvy @griffin_hector @FearTWD @MckeeveMichelle @LisaRSP @angelafabs Someth-aaang.. #FearTWD
Annotations	neg., neg., neg.	neu., pos., pos.	pos., neu., pos.
S_f	Negative	Neutral	Positive
C_f	0.87	0.58	0.67

In ID 13698, the image shows a dimly lit scene with individuals and cars, potentially evoking mixed interpretation, depending on the perspective of the viewer. Similarly, the text is conversational and includes hashtags related to entertainment and fandom, which could be perceived as positive or neutral depending on context and annotator bias. The framework aggregates these conflicting inputs to assign a positive sentiment, supported by moderate confidence ($C_f = 0.67$). This instance highlights how cultural, contextual, and personal biases can introduce ambiguity in both text and image sentiment annotations.

4.3 Comparative analysis

The text-based models consistently outperform image-based models in both crisp and fuzzy evaluations, due to the explicit sentiment cues inherent in textual data. This clarity allows the text models to achieve higher macro-F1 scores across all evaluation settings. For example, in the crisp evaluation, the text model achieves a macro-F1 score of 0.476, significantly surpassing the image-based model's 0.366. This performance disparity remains evident in the fuzzy evaluation, where the text model attains a macro-F1 score of 0.736 at $\delta = 0.3$, compared to the image-based model's 0.499 at the same threshold. These findings underscore the inherent limitations of image-only approaches, particularly in identifying neutral and negative classes, where visual cues tend to be less distinct or ambiguous.

Furthermore, while the application of ADASYN helped mitigate class imbalance, it is important to consider its potential drawbacks. Since this approach generates synthetic samples using k-nearest neighbours, it may introduce local interpolation artifacts, particularly in sparsely populated regions of the feature space. This effect can sometimes lead to the generation of synthetic samples in less representative areas, which could impact generalisation. However, despite a slight under-performance in the positive class, the overall gains, particularly in the minority negative class, justify its use. These results suggest that, while ADASYN is a useful technique for balancing the dataset, additional augmentation strategies could further enhance performance.

In contrast, multimodal models demonstrate superior performance by effectively leveraging the complementary strengths of text and image data. By assigning a higher weight to text features ($W_t = 1.5$) relative to image features ($W_i = 0.9$), the multimodal models capitalize on the clearer sentiment cues in text while using image data as a supplementary contextual source. This strategic weighting yields improved performance across metrics. For instance, in the crisp evaluation, the multimodal model achieves a macro-F1 score of 0.535, outperforming both text-only and image-only models. Notably, the improvement is most pronounced in the minority negative class, where the multimodal model reaches an F1 score of 0.382, compared to 0.16 for text-only and 0.087 for image-only models.

The benefits of multimodal integration are further amplified in fuzzy evaluation, which accommodates partial correctness in ambiguous instances. At $\delta = 0.3$, the multimodal model achieves a macro-F1 score of 0.683, showcasing its ability to handle low-confidence cases effectively. Class-specific improvements are evident as well, the negative class improves to an F1 score of 0.493, the neutral class rises to 0.681, and the positive class maintains a high score of 0.832. These results highlight the potential of fuzzy evaluation to enhance performance in challenging scenarios, such as those involving conflicting or ambiguous annotations.

Weighted evaluation metrics, particularly the weighted-F1 score, offer additional insights by emphasizing high-confidence instances. This approach provides a realistic representation of model reliability in handling clear and unambiguous cases. For instance, under fuzzy evaluation

at $\delta = 0.3$, the weighted-F1 score of the text model increases to 0.884, compared to its macro unweighted counterpart of 0.736. However, it is important to note that weighted metrics are not directly comparable across modalities, as confidence levels vary significantly between text, image, and multimodal data. This limitation is especially relevant for multimodal models, where the aggregated confidence for tweets is often reduced due to inter-modality disagreements.

In this work, a modular architecture has been selected, which uses Xception to extract visual features and DeBERTa to process textual information, instead of integrated multimodal models such as CLIP [35] or BLIP [36] which jointly encode image and text, however, the joint representation is highly coupled and can act as a 'black box'. Also, the proposed model offers representations that can be mapped more naturally to fuzzy sets and linguistic rules, something that is less straightforward in models whose output is a dense embedding that is difficult to map transparently. Therefore, this choice is motivated by several factors: it allows for greater control and interpretability of each modality, enables seamless integration with fuzzy logic systems, and offers flexibility to adapt each model to the specific needs of the task. In addition, by keeping image and text encoders separate, it is easier to analyse and understand the individual contribution of each type of data to the final decision. Regarding visual sentiment analysis, there are more advanced vision models, such as Vision Transformers (ViTs) [37] or Swin Transformers [38], which are able to capture complex contextual relationships within images. However, these models present a high computational cost and a large set of images to train. Finally, the proposed approach is more computationally efficient, as it does not require large datasets in which image and text are paired, which is necessary for training or tuning embedded models.

4.4 Ablation and component contribution analysis

Table 12 reports an incremental ablation under crisp evaluation for each modality.

Table 12 Incremental ablation under crisp evaluation

Modality	Component	macro-F1	MCC	BA	Kappa
Text	Baseline	0.476	0.301	46.40%	0.295
Text	Metadata	0.485 (+0.009)	0.311 (+0.010)	47.10% (+0.70pp)	0.305 (+0.010)
Text	CW	0.499 (+0.014)	0.316 (+0.005)	48.00% (+0.90pp)	0.311 (+0.006)
Text	ADASYN	**0.566** (+0.081)	**0.339** (+0.028)	**56.70%** (+9.60pp)	**0.339** (+0.034)
Image	Baseline	0.366	0.123	38.03%	0.107
Image	Colour	0.411 (+0.045)	0.163 (+0.040)	40.69% (+2.66pp)	0.159 (+0.052)
Image	CW	0.416 (+0.005)	0.167 (+0.004)	41.04% (+0.35pp)	0.163 (+0.004)
Image	ADASYN	**0.421** (+0.010)	**0.170** (+0.007)	**43.47%** (+2.78pp)	**0.169** (+0.010)
Multi-modal	Baseline	0.535	0.298	51.60%	0.289
Multi-modal	CW	0.553 (+0.018)	0.309 (+0.011)	53.00% (+1.40pp)	0.308 (+0.019)
Multi-modal	ADASYN	**0.574** (+0.039)	**0.330** (+0.032)	**59.40%** (+7.80pp)	**0.329** (+0.040)

As summarised in Table 12, augmenting the textual stream with metadata increases macro-F1 by 0.009 and balanced accuracy by 0.70pp, while adding colour cues to the visual stream improves macro-F1 by 0.045 and balanced accuracy by 2.66pp. Building on these enriched baselines, the class-imbalance correction introduced by ADASYN delivers the largest marginal improvements, culminating in the multimodal model with +0.039 macro-F1 and +7.80pp balanced accuracy, whereas the confidence-weighted loss contributes smaller yet systematic increments across modalities (e.g., +0.018 macro-F1 and +1.40pp balanced accuracy in the multimodal setting). Taken together, the progression from unimodal to enriched inputs, and then to rebalanced training, indicates that most of the performance lift is explained by distributional rebalancing, with confidence weighting acting as a complementary regulariser that consolidates the gains.

4.5 Feasibility analysis and applications

The previous section has analysed and compared the different approaches proposed. In this section, the model will be studied from a computational efficiency and complexity point of view. It will also highlight possible applications of the model beyond the problem presented in this paper, as well as the weaknesses and strengths of the model.

From the point of view of computational efficiency and complexity, the proposed models are complex and have a high computational cost during their training phases. However, these aspects, which may seem negative, do not affect the performance of the application when deployed in a real online system. This is because the high complexity and computational time are primarily used in the construction phase, whereas in the inference phase, the model remains highly efficient, with response times well below one second. The inference pipeline is optimised by performing the extraction of features from text using word embeddings (~ 0.005 s), image processing with Xception (~ 0.022 s), and colour feature extraction (~ 0.003 s) in parallel, minimizing latency. These features are then sequentially processed by DNN, which completes its inference in approximately $0.002 - 0.005$ s, ensuring rapid predictions. The system runs on a 16-core Intel(R) Xeon(R) Silver 4216 processor, 384 GB of RAM, and two NVIDIA A100 GPUs with 40 GB of memory each. Moreover, considering that these systems need to be retrained with new information, this would not be a problem either. During inference, the system would continue its normal operation, while offline it would retrain the model, and once ready, the model replacement would be automatic, ensuring uninterrupted service. Furthermore, it is worth noting that the system was designed and evaluated in a high-performance computing environment. This context implies that the primary focus of our work has been operational efficiency rather than optimisation for highly constrained edge devices. Nevertheless, the inference stage remains extremely lightweight, with total response times under 30 milliseconds. This responsiveness makes the system well-suited for real-time sentiment applications. Regarding the additional components introduced in this work, the computational footprint of ADASYN and the fuzzy evaluation layer is modest. ADASYN is executed only once on the training split to generate synthetic minority instances, so its cost is amortised over the offline training process and does not affect online latency. It increases the number of training samples during optimisation, but the inference phase relies on the same model size and feature dimensionality as the baseline, so memory usage at deployment time remains unchanged. The fuzzy evaluation layer operates on top of the three class probabilities and the tweet-level confidence with a

small number of arithmetic operations per instance, which adds no measurable overhead to the inference times reported above when compared with the forward passes of DeBERTa, Xception and the DNN. Although energy consumption and memory usage were not explicitly measured, the system architecture, which relies on pre-extracted features, parallel processing of modalities, and non-fine-tuned models, indicates a reasonable memory footprint. In addition, the model can be retrained offline without disrupting inference, allowing continuous service even as data conditions evolve. This parallelism in training and modular architecture ensures consistency across modalities and supports efficient updates without service interruption. Therefore, we can state that the system is computationally efficient and that the complexity of the model does not prevent real-time deployment in typical social media monitoring scenarios.

The model proposed in this paper has been evaluated using a dataset of tweets. However, its application goes beyond the world of social networks. From a business perspective, marketing is one of the most important aspects of success [39]. This model can be applied to assess the sentiments conveyed in a proposed poster, advertisement, post, or similar. This could help to avoid conflicts in marketing, as advertisers may have a positive approach and yet the advertisement may not be conveying that feeling. In a health application, it may be even more useful to evaluate every single advertisement or post that is published. The health campaigns that governments are running to prevent diseases and/or breakdowns in health systems should resonate deeply with people [40]. Therefore, the more visions and points of view that people have of an advertisement, both its text and its image, the more influential the advertisement can become. The proposed model may also be ideal for this case, since by providing a fuzzy focus, a greater amount of information is provided to users. Thus, the model could provide the final fuzzy output, but it would also be feasible to provide intermediate outputs, thus providing a feasible amount of information for end users to decide with greater guarantees. Beyond these marketing and public-health scenarios, the same fuzzy, confidence-aware outputs and intermediate explanations make the approach suitable for intelligent environments such as social sensing dashboards, safety monitoring systems and customer-care automation interfaces, where risk-aware decision policies can benefit from calibrated sentiment estimates and sample-level confidence profiles.

A final point to note is the strengths and weaknesses of the proposed method. Among the strengths of the method, we find a complex but efficient model which combines a large amount of information and uses fuzzy logic to provide a more complete view of the information. In addition, the applicability of the proposal is wide and allows working in a transparent way to the user, since with an accessible and simple graphical interface any user could upload image and text and analyse the feelings that transmit together such information. Finally, we focus on the weaknesses of the model. As we have seen in the results, minority classes benefit little in the results. To alleviate this disadvantage, we need to study in a future line the possibility of creating synthetic data that are representative of the minority class in order to create a better balance of the classes. In addition, another possible line of work would be to look for new data sources and select the minority classes and use a data condensation for the majority classes to select those instances that are more representative. However, it is important to note that this negative aspect of minority classes is a problem that needs to be addressed in many everyday problems.

5 Conclusions and future work

This study has presented a modular framework for multimodal sentiment analysis that couples DeBERTa for text and Xception for images with lightweight auxiliary channels (metadata and dominant colour features) and a fuzzy-logic layer that aggregates annotator confidence and evaluates with a δ-tolerant criterion. The design targets real-world conditions where ambiguity and imbalance are the rule rather than the exception. On a markedly imbalanced X corpus, the multimodal model attains weighted-F1 of 0.493 for the negative class, 0.681 for the neutral class and 0.832 for the positive class, while an incremental ablation shows that auxiliary channels reinforce unimodal branches, ADASYN accounts for the largest gains, and confidence-weighted loss provides consistent secondary improvements without altering backbone capacity.

A second contribution we propose the selection of the fuzzy tolerance as a validation-time choice that maximises weighted-F1 while preserving agreement (WAI), and we document sensitivity curves across δ, with a stable operating point at $\delta^* = 0.3$ throughout modalities. Beyond three-way polarity, the same fuzzy layer extends naturally to fine-grained or partially overlapping taxonomies by aggregating annotator input into confidence-weighted soft targets and by awarding partial credit when predictions distribute probability across neighbouring labels. The framework remains backbone-agnostic and keeps modality separability, which improves interpretability and eases domain adaptation.

From the perspective of intelligent environments, such as social sensing dashboards, safety monitoring or customer-care automation, the proposed combination of confidence aggregation and δ-tolerant evaluation offers deployment-relevant properties: it provides calibrated, sample-level evidence that can be surfaced to decision policies; it exposes a controllable trade-off via δ so that tolerance can be tightened or relaxed according to risk; and it yields an agreement signal (WAI) that supports defer-or-act strategies. In practice, this enables risk-aware operation (e.g., deferring to a human when confidence weight or WAI fall below a threshold), targeted alerts when sustained negative sentiment emerges with high agreement, and auditable explanations because each prediction is linked to the contributing modality and confidence profile. Taken together, these elements advance research in intelligent environments and computational intelligence by showing how fuzzy logic driven confidence aggregation can be integrated with deep multimodal models at the levels of data labelling, class rebalancing and evaluation, which is essential for transparent and risk aware decision making in real deployments.

As future work, we will position the framework against recent vision and vision-language encoders, such as CLIP/ViT variants, BLIP-style models and multimodal transformers, under matched data and optimisation budgets, reporting both the crisp score (δ=0) and the fuzzy score at the validated operating point so that any differences reflect methodology rather than capacity. On the vision side, we will explore stronger backbones (ViT/Swin), region-focused cues that isolate sentiment-bearing elements, and richer semantics from facial, symbolic and contextual signals; on the data side, we will study augmentation strategies with greater semantic fidelity in both text (paraphrasing, BERT-based edits) and images (quality-constrained synthesis). For fine-grained sentiment, we plan to derive soft targets from a label-similarity kernel and, where overlap is pronounced, to consider class-specific tolerances δ_k, evaluating on datasets with nuanced categories (e.g., sarcasm). Finally, rather than discarding multimodal inconsistencies, we will investigate adaptive fusion and reinforcement learning to reconcile text–image evidence when the two streams diverge.

Author Contributions Conceptualization, Sara Balderas-Díaz (S.B.D.), Gabriel Guerrero-Contreras (G.G.C), Andrés Bueno-Crespo (A.B.C), Raquel Martínez-España (R.M.E); methodology, S.B.D., G.G.C, A.B.C, R.M.E; validation, S.B.D., G.G.C, A.B.C, R.M.E; formal analysis, S.B.D. and G.G.C; writing—review and editing, S.B.D., G.G.C, A.B.C, R.M.E.

Funding Funding for open access publishing: Universidad de Cádiz/CBUA. Financial support for this research has been provided under R+D+i AwESOMe Project PID2021-122215NB-C33, ONEFRE-4 PID2023-148104OB-C43, ONOFRE Project PID2020-112675RB-C44 and the ALLEGRO Project PID2020-112827GB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF A way to do Europe.

Data Availability Reproducibility materials, including code, configuration files, and scripts, are available at repository link. The MVSA-Multiple dataset is publicly accessible at Kaggle. Full access to the replication package is provided upon request from the authors, as the repository is shared with other ongoing research projects.

Declarations

Conflicts of Interest The authors declare that we have no conflict of interest.

Ethical Approval Not applicable.

References

1. Ghosh S, Ekbal A, Bhattacharyya P (2023) Natural language processing and sentiment analysis: perspectives from computational intelligence. Hybrid computational intelligence for pattern analysis and understanding, pp 17–47. Academic Press. https://doi.org/10.1016/B978-0-32-390535-0.00007-0
2. Das R, Singh TD (2023) Multimodal sentiment analysis: a survey of methods, trends, and challenges. ACM Comput Surv 55(13s):1–38. https://doi.org/10.1145/3586075
3. Gandhi A, Adhvaryu K, Poria S, Cambria E, Hussain A (2023) Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion 91:424–444. https://doi.org/10.1016/j.inffus.2022.09.025
4. Chandrasekaran G, Nguyen TN, Hemanth DJ (2021) Multimodal sentimental analysis for social media applications: A comprehensive review. Wiley Interdisc Rev Data Min Knowl Disc 11(5):1415. https://doi.org/10.1002/widm.1415
5. Ye J, Zhou J, Tian J, Wang R, Zhou J, Gui T, Zhang Q, Huang X (2022) Sentiment-aware multimodal pre-training for multimodal sentiment analysis. Knowl-Based Syst 258:110021. https://doi.org/10.1016/j.knosys.2022.110021
6. Hermida A, Mellado C (2020) Dimensions of social media logics: Mapping forms of journalistic norms and practices on twitter and instagram. Digit J 8(7):864–884. https://doi.org/10.1080/21670811.2020.1805779
7. Ortis A, Farinella GM, Battiato S (2020) Survey on visual sentiment analysis. IET Image Proc 14(8):1440–1456. https://doi.org/10.1049/iet-ipr.2019.1270
8. Zhu L, Zhu Z, Zhang C, Xu Y, Kong X (2023) Multimodal sentiment analysis based on fusion methods: A survey. Inf Fusion 95:306–325. https://doi.org/10.1016/j.inffus.2023.02.028
9. Hung BT, Thu NHM (2024) Novelty fused image and text models based on deep neural network and transformer for multimodal sentiment analysis. Multimed Tools Appl 83(25):66263–66281. https://doi.org/10.1007/s11042-023-18105-8
10. Yin Z, Du Y, Liu Y, Wang Y (2024) Multi-layer cross-modality attention fusion network for multimodal sentiment analysis. Multimed Tools Appl 83:60171–60187. https://doi.org/10.1007/s11042-023-17685-9

11. Ahuja G, Alaei A, Pal U (2024) A new multimodal sentiment analysis for images containing textual information. Multimed Tools Appl. https://doi.org/10.1007/s11042-024-19999-8
12. An J, Wan Zainon WMN (2023) Integrating color cues to improve multimodal sentiment analysis in social media. Eng Appl Artif Intell 126:106874. https://doi.org/10.1016/j.engappai.2023.106874
13. Salur MU (2022) Aydın: A soft voting ensemble learning-based approach for multimodal sentiment analysis. Neural Comput Appl 34:18391–18406. https://doi.org/10.1007/s00521-022-07451-7
14. Deng Y, Li Y, Xian S, Li L, Qiu H (2024) Mual: enhancing multimodal sentiment analysis with cross-modal attention and difference loss. Int J Multimed Inf Retriev 13. https://doi.org/10.1007/s13735-024-00340-w
15. Wang H, Ren C, Yu Z (2024) Multimodal sentiment analysis based on cross-instance graph neural networks. Appl Intell 54:3403–3416. https://doi.org/10.1007/s10489-024-05309-0
16. Paul A, Nayyar A et al (2024) A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis. Multimed Tools Appl 83(18):54249–54278. https://doi.org/10.1007/s11042-023-17601-1
17. Muñoz A, Martínez-España R, Guerrero-Contreras G, Balderas-Díaz S, Arcas-Túnez F, Bueno-Crespo A (2024) A multi-dl fuzzy approach to image recognition for a real-time traffic alert system. J Ambient Intell Smart Environ 1–17. https://doi.org/10.3233/AIS-230433
18. Zhi Y, Li J, Wang H, Chen J, Wei W (2024) A multimodal sentiment analysis method based on fuzzy attention fusion. IEEE Trans Fuzzy Syst. https://doi.org/10.1109/TFUZZ.2024.3434614
19. Nguyen TL, Kavuri S, Lee M (2019) A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. Neural Netw 118:208–219. https://doi.org/10.1016/j.neunet.2019.06.010
20. Wang X, Lyu J, Kim BG, Parameshachari BD, Li K, Li Q (2024) Exploring multimodal multiscale features for sentiment analysis using fuzzy-deep neural network learning. IEEE Trans Fuzzy Syst. https://doi.org/10.1109/TFUZZ.2024.3419140
21. Gutiérrez-Batista K, Vila MA, Martin-Bautista MJ (2021) Building a fuzzy sentiment dimension for multidimensional analysis in social networks. Appl Soft Comput 108. https://doi.org/10.1016/j.asoc.2021.107390
22. He H, Bai Y, Garcia EA, Li S (2008) Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE World Congress on Computational Intelligence), pp 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969
23. Niu T, Zhu S, Pang L, El Saddik A (2016) Sentiment analysis on multi-view social data. In: Tian Q, Sebe N, Qi G-J, Huet B, Hong R, Liu X (eds) MultiMedia modeling, pp 15–27. Springer, Cham. https://doi.org/10.1007/978-3-319-27674-8_2
24. Xu N, Mao W (2017) A residual merged neutral network for multimodal sentiment analysis. In: 2017 IEEE 2nd international conference on big data analysis (ICBDA), pp 6–10. https://doi.org/10.1109/ICBDA.2017.8078794
25. Guerrero-Contreras G, Balderas-Díaz S, Serrano-Fernández A, Muñoz A (2024) Enhancing sentiment analysis on social media: Integrating text and metadata for refined insights. In: 2024 International conference on intelligent environments (IE), pp 62–69. https://doi.org/10.1109/IE61493.2024.10599899
26. He P, Liu X, Gao J, Chen W (2021) Deberta: Decoding-enhanced bert with disentangled attention. In: International conference on learning representations. https://doi.org/10.48550/arXiv.2006.03654
27. Chersoni E, Santus E, Huang C-R, Lenci A (2021) Decoding word embeddings with brain-based semantic features. Comput Linguist 47:1–36. https://doi.org/10.1162/coli_a_00412
28. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258. https://doi.org/10.1109/CVPR.2017.195
29. Lin M, Chen Q, Yan S (2014) Network in network. https://doi.org/10.48550/arXiv.1312.4400
30. Han J, Lee Y (2024) Image sentiment considering color palette recommendations based on influence scores for image advertisement. Electron Commerce Res 1–28. https://doi.org/10.1007/s10660-024-09851-4
31. Ruan S, Zhang K, Wu L, Xu T, Liu Q, Chen E (2021) Color enhanced cross correlation net for image sentiment analysis. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2021.3118208
32. Lin C, Mottaghi S, Shams L (2024) The effects of color and saturation on the enjoyment of real-life images. Psychon Bull Rev 31(1):361–372. https://doi.org/10.3758/s13423-023-02357-4
33. Zhao Z, Zhu H, Xue Z, Liu Z, Tian J, Chua MCH, Liu M (2019) An image-text consistency driven multimodal sentiment analysis approach for social media. Inf Process Manag 56(6):102097. https://doi.org/10.1016/j.ipm.2019.102097
34. Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image-text sentiment analysis via deep multimodal attentive fusion. Knowl Based Syst 167:26–37. https://doi.org/10.1016/j.knosys.2019.01.019
35. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp 8748–8763. PMLR

36. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning, pp 12888–12900. PMLR
37. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
38. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
39. Movahed AB, Movahed AB, Nozari H (2024) Opportunities and challenges of marketing 5.0. Smart Sustain Interact Market 1–21
40. Thapliyal K, Thapliyal M, Thapliyal D (2024) Social media and health communication: A review of advantages, challenges, and best practices. Emerg Technol Health Liter Med Pract 364–384

Authors and Affiliations

Sara Balderas-Díaz[1] · Gabriel Guerrero-Contreras[1] · Andrés Bueno-Crespo[2] · Raquel Martínez-España[3]

✉ Sara Balderas-Díaz
sara.balderas@uca.es

Gabriel Guerrero-Contreras
gabriel.guerrero@uca.es

Andrés Bueno-Crespo
abueno@ucam.edu

Raquel Martínez-España
raquel.m.e@um.es

[1] Department of Computer Science and Engineering, University of Cadiz, Av. Universidad de Cádiz, 10, Puerto Real 11519, Cádiz, Spain

[2] Escuela Politécnica Superior, Universidad Católica de Murcia (UCAM), Campus de Murcia, Guadalupe 30107, Murcia, Spain

[3] Department of Information and Communications Engineering, University of Murcia, Computer Faculty, Campus de Espinardo, Espinardo 30100, Murcia, Spain