NELSON ANDRÉ SARDO DA SILVA

MULTIMODAL SENTIMENT CLASSIFIER FOR VARIOUS ENVIRONMENTS CONTEXTS

UNIVERSIDADE DO ALGARVE
Instituto Superior de Engenharia
2024

NELSON ANDRÉ SARDO DA SILVA

MULTIMODAL SENTIMENT CLASSIFIER FOR VARIOUS ENVIRONMENTS CONTEXTS

Master´s Degree in

Electrical and Computer Engineering

(Specialization in Information Technology and Telecommunications)

Work carried out under the guidance of:

Professor Ph.D. Pedro J. S. Cardoso

Professor Ph.D. João M. F. Rodrigues

UNIVERSIDADE DO ALGARVE
Instituto Superior de Engenharia
2024

MULTIMODAL SENTIMENT CLASSIFIER FOR VARIOUS ENVIRONMENTS CONTEXTS

Declaration of authorship of the work

I hereby declare to be the author of this work, which is original and unpublished. Authors and works consulted are properly cited in the text and included in the reference list.

_____

(Nelson André Sardo da Silva)

ACKNOWLEDGEMENTS

Firstly, my sincere thanks to my supervisors for giving me the opportunity to carry out this thesis in the Human-Centred AI, Sentiment Analysis, and Affective Computing areas, that I have always been interested in learning and researching. I couldn't have asked for better supervisors than Professor João Rodrigues and Professor Pedro Cardoso, whose knowledge and experience guided me every step of the way. Without their constant support, monitoring, openness, and availability, my performance and development would not have been the same.

I would also like to thank the University of Algarve, particularly the ISE Visual Computing Lab, for providing me with a computer to carry out some exhaustive tasks that were essential for the development of the dissertation.

I'd like to acknowledge the emotional support provided by my family, especially my mother who encouraged me in the most difficult times to not give up.

Finally, I would like to express my special thanks to all the teachers who contributed to my academic training over the years, for all their availability and knowledge transmitted.

To all the nominees, I would like to express my recognition and sincere praise.

Thank you.

RESUMO

A análise de sentimentos é um método eficaz para determinar a opinião pública. As publicações nas redes sociais têm sido objeto de muita investigação, principalmente devido à enorme e diversificada base de utilizadores dessas plataformas que partilham regularmente opiniões sobre praticamente todos os assuntos. No entanto, nas publicações (*posts*) compostas por um par texto-imagem, a descrição escrita pode ou não transmitir o mesmo sentimento que a imagem. Este estudo utiliza modelos de aprendizagem automática para a avaliação automática do sentimento de pares de texto e imagem(ns). Os sentimentos derivados da imagem e do texto são avaliados de forma independente e associados (ou não) para formar o sentimento global, devolvendo o sentimento da publicação e a discrepância entre os sentimentos representados pelo par texto-imagem. A classificação do sentimento da imagem é dividida em 4 categorias: "interior" (IND), "exterior feito pelo homem" (OMM), "exterior não feito pelo homem" (ONMM) e "interior/exterior com pessoas em segundo plano" (IOwPB). No final, os resultados são consolidados num modelo de classificação do sentimento da imagem (ISC), que pode ser comparado com um classificador holístico do sentimento da imagem (HISC), mostrando que o ISC obtém melhores resultados do que o HISC. Para um subconjunto de dados do Flickr, a classificação do sentimento das imagens, por categoria, atingiu uma exatidão de 68,50% para IND, 83,20% para OMM, 84,50% para ONMM, 84,80% para IOwPB e 76,45% para ISC, em comparação com 65,97% do HISC. Para a classificação do sentimento do texto, num subconjunto da base de dados B-T4SA, foi alcançada uma exatidão de 92,10%. Por fim, a combinação texto-imagem, num conjunto de dados privado, obteve uma exatidão de 78,84%.

Palavras-Chave: Análise de Sentimentos; Computação Afetiva; Inteligência Artificial Centrada no Humano; Classificador de Sentimentos Multimodal.

ABSTRACT

Sentiment analysis is an effective method for determining public opinion. Social media posts have been the subject of much research, due to the platforms' enormous and diversified user base that regularly share thoughts on nearly any subject. However, on posts composed by a text-image pair the written description may or may not convey the same sentiment as the image. The present study uses machine learning models for the automatic sentiment evaluation of pairs of text and image(s). The sentiments derived from the image and text are evaluated independently and merged (or not) to form the overall sentiment, returning the sentiment of the post and the discrepancy between the sentiments represented by the text-image pair. The image sentiment classification is divided into 4 categories: "indoor" (IND), "man-made outdoors" (OMM), "non-man-made outdoors" (ONMM), and "indoor/outdoor with persons in the background" (IOwPB). The results are then ensembled into an image sentiment classification model (ISC), that can be compared with a holistic image sentiment classifier (HISC), showing that the ISC achieves better results than the HISC. For the Flickr sub-dataset, the sentiment classification of images achieved an accuracy of 68.50% for IND, 83.20% for OMM, 84.50% for ONMM, 84.80% for IOwPB, and 76.45% for ISC, compared to the 65.97% for HISC. For the text sentiment classification, in a sub-dataset of B-T4SA, an accuracy of 92.10% was achieved. Finally, the text-image combination, in a private dataset, achieved an accuracy of 78.84%.

Keywords: Sentiment Analysis; Affective Computing; Human-Centered AI; Multimodal Sentiment Classifier.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

ACRONYM	*Meaning*
AffC	Affective Computing
AI	Artificial Intelligence
BLIP	Bootstrapping Language-Image Pre-training
BOW	Bag of Words
B-T4SA	Twitter for Sentiment Analysis sub-dataset
CLIP	Contrastive Language-Image Pre-training
CMFeed	Controllable Multimodal Feedback Synthesis
dis.	Discrepancy
DL	Deep Learning
ECGs	Electrocardiograms
EEGs	Electroencephalographs
GPT-3	Generative Pre-trained Transformer 3
GPU	Graphics Processing Unit
HCAI	Human-Centered Artificial Intelligence
HISC	Holistic Image Sentiment Classifier
HTML	HyperText Markup Language
IND	Indoor
IoT	Internet of Things
IOwPB	Indoor/Outdoor with persons in the background
ISC	Image Sentiment Classifier / Classification
ISP	Image Sentiment Polarity
LLM	Large Language Model
LSCm	Landscape Sentiment Classification model
ML	Machine Learning
MSC	Multimodal Sentiment Classifier

MVSO	Multilingual Visual Sentiment Ontology
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Neutral Network
OMM	Man-Made Outdoors
ONMM	Non-Man-Made Outdoors
RAM	Random Access Memory
R-CNN	Region-based Convolutional Neural Network
RF	Random Forest
SCv	Sentiment Classifier Vector
SGD	Stochastic Gradient Descent
SIS	Simula Image Sentiment
T4SA	Twitter for Sentiment Analysis
TSC	Text Sentiment Classification
VGG	Visual Geometry Group
VLPCA	Vision-Language Pre-Training Model based on Cross-Attention

1

INTRODUCTION

The analysis of sentiments and emotions is an area that has been growing recently due to the evolution of Artificial Intelligence (AI). In particular, the branches of Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) have made it possible to extract relevant information from multimedia data such as texts, images, videos, and sound, among other types of sources. In addition, the exponential growth of social networks in recent years has generated a large amount of textual and visual data essential for sentiments and emotions analysis. Since sentiments and emotions are directly related and allow us to understand the users' emotional state, analysing them is important for companies and governments to understand people's satisfaction with a product and marketing campaign or their reaction to political, climatic, and other issues.

Nowadays, Society 4.0 [1] is defined as a society that prioritizes the interests of its citizens. It is a concept that has been developed to transform regions into the foundation of a future citizen-centric society. This transformation is enabled by several technological developments that allow for the decentral production of most primary human necessities, such as energy, food, health, and education. Following this, Society 4.0 has improved the interaction between humans and machines, transforming standard industrial factories into smart factories through the implementation of technologies such as the Internet of Things (IoT), big data analysis, 3D printing, robotics, and AI, resulting in an improvement in the level of production, time management, relations with suppliers and customers, and other aspects. However, to improve the population's quality of life, focusing even more on the citizens, a proposal emerged in Japan for the creation of Society 5.0 [2]. In the Society 5.0 the main focus is not exactly on industry or the economy, but on society, especially the current urgent social demands to solve

problems in cities such as the ageing of the population, improvements in health services, infrastructure, mobility in cities and between cities, natural disasters, climate changes, among others, with the help of technology, making it even more present and essential in human life. All this emphasises the importance of people's sentimental analysis of current environmental and social issues.

In contrast to the numerous studies focused on analysing emotions and sentiments through people's facial expressions and poses [3, 4], this study focuses on analysing text-image posts, where the images depict environments and humans are not the primary source of analysis. This approach allows for a better understanding and enhancement of posts related to landscapes, natural disasters, environmental preservation, climate change, etc.

In text-image posts, it is "relatively" easy to analyse the emotion or the sentiment that a text expresses. But usually, the texts come associated with one or more images, and if those images do not have humans in the foreground, in a frontal position, there is always a question if the image reinforces the emotion or sentiment of the text or if it is only to frame it. In this context a multimodal approach needs to be used to analyse the text, the image, and their combination, in order to detect the sentiment associated with the post. In addition to the multimodal approach, the ensemble technique can also be used, which enhances the robustness, generalization, and accuracy of sentiment analysis by combining the results of the individual models.

As a result, it is not necessary to fully train the individual models, since the ensemble technique and the multimodal approach use the results of other trained models. In consequence, the energy consumed to train them is significantly reduced, contributing to the reduction of the algorithm/model's "carbon footprint". Nevertheless, training models can consume a lot of energy, so it is important to use high-quality training data to avoid inaccuracies, high-quality baseline algorithms, tuning, and other efficient configuration strategies. More details on the implementation and operation of the multimodal approach and ensemble technique adopted for this thesis will be explained in the following chapters.

In summary, this dissertation focuses on developing a multimodal framework that can analyse and classify the sentiment of a text-image pair, also called a social media post, in which the human being is not the images' primary focus. This approach allows to understand the real (human) sentiments and emotions when writing the post, and the relation of the text with the image associated, especially when images are natural landscapes, cities, and the interiors of buildings and structures. Furthermore, the focus of this thesis is also to contribute

to the social development envisioned by Society 5.0. This includes providing insights through the sentimental analysis on environmental, social and other issues related, as well as helping to promote a sustainable technological future in which the aim is to increasingly reduce the energy used by machines to do this type of analysis. Consequently, all this increases and improves the human-machine relationship.

1.1 MAIN OBJECTIVES

The main objective of this dissertation is to develop a multimodal sentiment classification model that analysis text and image from posts, specially focusing on images set in indoor or outdoor environments. The specific sub-objectives include:

a) Analyse and compile the state-of-the-art and existing datasets;

b) Develop and/or adapt a set of models for text sentiment analysis;

c) Develop and/or adapt a set of models for image sentiment analysis;

d) Develop an ensemble model for text sentiment analysis;

e) Develop an ensemble model for image sentiment analysis;

f) Develop an ensemble (multimodal) model for text & image sentiment analysis;

g) Develop an indicator to compute de discrepancy between the sentiments expressed by the text and image when both are used for the same purpose (e.g., same post);

h) Test the models using text and images from posts featuring different environments, such as indoor man-made contexts, outdoor man-made contexts, and indoor/outdoor posts with people in the background;

i) Write at least one scientific article;

j) Write the dissertation as a compilation of scientific articles.

1.2 CONTRIBUTIONS

The main contributions of this dissertation include:

(i) An image classification sentiment model that works in different scenes/environments;

(ii) A framework that combines image and text sentiment classification, returning the multimodal sentiment classification attached with the discrepancy (text and image sentiment) metric.

Two scientific articles were developed, namely:

a. *Silva, N., Cardoso, P. J. S., & Rodrigues, J. M. F. (2024). Sentiment Classification Model for Landscapes,* In: Antona, M., Stephanidis, C. (eds), Universal Access in Human-Computer Interaction. HCII 2024. Lecture Notes in Computer Science, vol. xx Springer, Cham. DOI: xx (waiting for the final proceedings).
b. *Silva, N., Cardoso, P. J. S., & Rodrigues, J. M. F. (2024). Multimodal Sentiment Classifier Framework for Different Scene Contexts*. Appl. Sci. 2024, 14, 7065. DOI: https://doi.org/10.3390/app14167065

1.3 ORGANIZATION OF THE REPORT

The report of this dissertation consists in a compilation of two publications, one accepted at the 26th International Conference on Human-Computer Interaction, held in Washington DC, USA, on 29 June - 4 July 2024, and one published in the Applied Sciences Journal, in August 2024.

Both publications serve as chapters of the dissertation, being some sections very similar, due to the works continuity. However, after consulting with the dissertation supervisors, it was decided to retain the full text of both publications to enhance the readability of each chapter. Chapter 2, the first paper, focuses on automatic sentimental evaluations for outdoor environments, in particular natural environments. This research can provide valuable insight for applications like tourism and marketing planning because it helps to understand how people perceive, appreciate, and engage with their surroundings. Often, these places are portrayed in social media posts accompanied by a written description that may or may not convey the same emotion as the image. In this paper, we study the automatic sentimental evaluations of pairs of photos and texts, where the depicted images are from natural environments – landscapes. During the analysis, the sentiments derived from the image and text are evaluated independently. These individual sentiments are then merged to form the overall sentiment associated with the text-image pair. The analysis provides the sentiment (positive, negative, or neutral) associated with the image, the text, the combination of the image and the text, and the discrepancy between the sentiments represented in the two. Overall, an ensemble of deep-learning models was used for images classification, and an ensemble of machine-learning models for the text and for the combination of image-text, the latter applied if the discrepancy justifies that ensemble. According to preliminary findings, for

the landscape photo-text combination, in our private dataset, we achieved an accuracy of 78.75%.

Chapter 3, the second paper, also focuses on sentiment analysis, which is an effective method for determining public opinion. As before, this chapter's concept is supported on social media posts which have been the subject of much research, due to the social media platform's enormous and diversified user base, that regularly share thoughts on nearly any subject. However, on posts composed by a text-image pair the written description may or may not convey the same sentiment as the image. The study uses machine learning models for the automatic sentiment evaluation of pairs of text and image(s). The sentiments derived from the image and text are evaluated independently and merged (or not) to form the overall sentiment, returning the sentiment of the post and the discrepancy between the sentiments represented by text-image pair. The image sentiment classification is divided into 4 categories: "indoor" (IND), "man-made outdoors" (OMM), "non-man-made outdoors" (ONMM), and "indoor/outdoor with persons in the background" (IOwPB), and then ensembled into an image sentiment classifier (ISC), that can be compared with a holistic image sentiment classifier (HISC). Results show that the ISC achieves better results than the HISC. For the Flickr sub-dataset, the sentiment classification of images achieved an accuracy of 68.50% for IND, 83.20% for OMM, 84.50% for ONMM, 84.80% for IOwPB, and 76.45% for ISC, compared with 65.97% for HISC. For the text sentiment classification, in a sub-dataset of Twitter for Sentiment Analysis (B-T4SA), an accuracy of 92.10% was achieved. Finally, the text-image combination, in the authors' private dataset, achieved an accuracy of 78.84%.

As noted, both chapters/papers contain sub-sections with similarities, including the introduction and the state-of-the-art sections. As mentioned above, the first paper focuses only on landscapes (pair text-image, where images are landscapes), and the second on generic environment analysis. The methods of the second paper are based (as explained in the text) on the method presented in the first paper. The solution for the discrepancy metric evolved from the first to the second paper. The final conclusions and future work of the dissertation are the compilation of the conclusions of both papers.

In summary, the present chapter focuses on the introduction, objectives and contribution of the dissertation, Sections 2.1 and 3.1 focus on the introduction of the respective papers, Sections 2.2 and 3.2 do contextualization and state-of-the-art, Sections 2.3 and 3.3 the datasets and sub-datasets used, Sections 2.4 and 3.4 the methods, followed by Sections 2.5 and 3.5 of tests and results, and finalizing with Sections 2.6 and 3.6 with the conclusions and future work.

The last chapter, Chapter 4, will summarize our final conclusions and show some future work.

2

SENTIMENT CLASSIFICATION MODEL FOR LANDSCAPES

2.1 INTRODUCTION

The goal of Human-Centered Artificial Intelligence (HCAI) is to create technologies that assist people in carrying out various daily tasks, while also advancing human values, such as rights, fairness, and dignity [5]. By promoting human's autonomy, well-being, and control over future technologies, HCAI also seeks to strike a balance between human control and (full) automation, as an interdisciplinary field combining computer science, psychology, and neuroscience. In a related field, Affective Computing (AffC) integrates the disciplines of emotion recognition and sentiment analysis, being supported by various types of physical information, such as text, audio (speech), visual data (e.g., facial expression, body posture, or environment), or physiological signals (e.g., EEGs – electroencephalography or ECGs – electrocardiograms). Within this framework, AffC can be built on either unimodal or multimodal data [6].

The applications of HCAI and AffC are numerous. For instance, a machine, in *lato sensu*, should be developed to collaborate or learn to work with humans, like an interpersonal connection. However, feelings and sentiments are essential to interpersonal connections, and those must be integrated into any machine that interacts with humans.

On the other hand, social media platforms have a growing importance today, influencing people to visit places, search and buy products, change lifestyles, change their way of thinking

about a subject etc. Within the volume of daily posts on many of those platforms, the necessity and opportunity to analyse and understand the emotion and sentiment that those messages carry have emerged. Video, photos, and short texts are the fastest and simplest type of publication to attract users to click, buy, and read about a particular subject or product, which is why social networks, such as Instagram or X (previously Twitter), become quite popular in recent years.

In this context, significant developments have been made in sentiment detection through text and image (photo) analysis. Even though sometimes the image transmits a different sentiment from the text, image and text can be used to complement each other in the analysis of sentiment in a post, being this information important for marketers, and in the development of machines (e.g., robots and interfaces) that follows the HCAI principles.

In this chapter, we focus on classifying sentiments in natural environments, i.e., landscape posts supported by photo-text data, a multimodal approach. For example, this classification problem is completely different from the detection of emotions and sentiments from facial expressions, speech, or body posture [7]. Sentiment analysis becomes more intricate when we segment these settings into indoor and outdoor spaces, which probably should be dealt with separately to attain more accurate methods. Further, in making this differentiation, there is the aspect that indoor spaces are typically man-made and outdoor environments are split into urban (man-made) and natural (landscapes) environments. For the types of problems previously presented, but even with more emphasis in the latter case (namely, landscapes sentiment analysis), color, texture, edges, line type, and orientation play a key role in the attraction and sentiment that the photo carries. In fact, when trying to extract sentiments from natural environments, color data may be one of the most important features at play [8].

In summary, this work presents a model for a landscape sentiment classifier based on image (photo) and associated text, frequently observed in posts on social media platforms, the Landscape Sentiment Classification model (LSCm). The LSCm aggregates image and text information, returning the sentiment classification and discrepancy between the image and text-predicted sentiments.

The main contribution of this work is three-fold: (i) classify sentiments in natural scene images (landscapes), (ii) classify the sentiments linked with the text associated with a landscape, and (iii) combine image and text classification, returning the information about its discrepancy.

The present section introduces the goal of the work as well as the list of main contributions,

Section 2.2 presents the contextualization and state of the art, Section 2.3 introduces the datasets used, Section 2.4 details the model, and Section 2.5 outlines the tests, results and respective discussion. Finally, Section 2.6 presents the conclusions and future work.

2.2 CONTEXTUALIZATION AND STATE-OF-THE-ART

The term interpersonal relationship refers to the association, warmth, friendliness, and dominance between two or more people, expressed when relationships are formed, reciprocated, or deepened [9]. Simultaneously, impersonal human-machine interactions hinder broader communication, making difficult the reciprocal or close ties between humans and machines (devices and/or interfaces). In this context, automatic emotion and sentiment analysis methods are the automated processes of analysing information to estimate the emotion [10] (e.g., categorized as happiness, sadness, fear, surprise, disgust, anger, and neutral) or sentiment [11, 12] (typically limited to positive, negative, and neutral). Sentiment classification has been a popular research topic in recent years, and significant progress has been made. Here we should notice that sentiments and emotions are interdependent, nevertheless are different concepts [12, 13]. The sentiment is a mental attitude related to positive, negative, or neutral evaluation/thought of something, and can be influenced by factors such as emotion, past experiences, cultural background, personal beliefs, age, or even gender [13]. The categorization of user emotions and sentiments goes beyond the traditional approach of identifying facial expressions, which focuses on alterations in the facial features and actions of a subject.

However, the examination of what feeling or attitude is evoked by a landscape, which is an umbrella term for the geographic features that mark or are typical of a certain location, and is sometimes referred to as a picture of natural scenery, is far more complex. It is therefore important to think about the category and context of such images while analysing them. An image of a beach or the ocean will often have blue as the predominant color, which is typically connected to serenity, while an image of a forest will emphasize green tones, which are typically connected to harmony. Color should not be the single feature to analyse, as the meaning of color might vary depending on where it is used; for instance, red can in some cases signify rage, love, or frustration [8, 14, 15, 16]. In addition, just like for music, different people may emotionally interpret color dissimilarly.

There are several types of emotion classification and different authors divide them into

several levels/sublevels [13]. As mentioned before, the six basic emotions (usually complemented with the neutral emotion) was a classification done by the famous psychologist Paul Ekman [10], based on universal facial expressions, but is not the focus of this work. Other authors also proposed alternative classifications, like Robert Plutchik [17] that defined eight basic/primary emotions, based on adaptative biological processes, namely: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Plutchik developed a color wheel, called Plutchik's wheel, to represent these emotions, with each emotion being associated with a specific color. In this case, emotions can be classified into different intensities and combinations, i.e., the primary emotions are then combined in different ways to create secondary and tertiary emotions, which are represented by different shades and hues of colors, totalling 24 emotions.

The emotions shown in Plutchik's wheel are divided into two main sentiments: positive and negative. The positive sentiment emotions are joy, trust, anticipation, and surprise, while the sadness, disgust, fear, and anger emotions are considered negative. However, the classification of emotions into positive and negative sentiments can be somewhat subjective and may depend on individual and cultural factors, as already mentioned [16]. The sentiment is often longer and more stable than emotion, which can change rapidly in response to changing stimuli and contexts [18].

Several authors worked in sentiment classification. Ortis et al. [13], in 2019, presented an overview of image sentiment analysis. The authors discussed the major issues and outlined opportunities and challenges in the area. Gaspar and Alexandre [19] presented a multimodal sentiment analysis (image and text) model for the classification of the content of composite comments in social media. The method is divided into three main parts: a text analysis, an image classifier, and a method that analyses the class content of an image, checking the probability that it belongs to one of the possible classes. They worked with the Twitter for Sentiment Analysis (T4SA) dataset [20], which has three million tweets (images and text) classified into three sentiments: positive, negative, and neutral. Oliveira et al. [21] presented the OutdoorSent, a framework to classify the sentiment of generic outdoor images shared by users on social networks. The authors also evaluate how the merging of deep features and semantic information derived from the scene attributes can improve classification and cross-dataset generalization performance. The same authors also propose a dataset of geolocalized urban outdoor images extracted from Flickr and labelled (by at least five volunteers) the samples as positive, negative, or neutral.

A color cross-correlation net for image sentiment analysis was presented in [8]. The architecture not only leverages contents and colors simultaneously, but also considers their correlations. The authors used a pre-trained convolutional neural network to extract content features and color moments to collect color features from multiple color spaces. Then, they propose a cross-correlation method to model the relationships between content and color features, with an attention mechanism and sequence convolution, in which sentiment expressing of content and color can be enhanced by each other, integrating these two types of information for better results in the end.

Chatzistavros et al. [22] presented a deep-learning architecture for sentiment analysis on 2D images of urban and indoor spaces. In [23] is analysed the sentiment from disaster images in social media. In [24], based on a gated attention mechanism, a multimodal (text and image) sentiment classification model is presented, where the image feature is used to emphasize the text segment by the attention mechanism, allowing the model to focus on the text that affects the sentiment polarity. Moreover, the gating mechanism enables the model to retain useful image information, while ignoring the noise introduced during the fusion of image and text.

It is important to stress, that there are a huge number of models available to do text analysis [12], which usually follow the following steps:

(i) *Text processing*: The text is cleaned to increase the sentiment prediction accuracy using techniques such as tokenization, stop word removal, stemming, lemmatization, emote and emoji conversion, and removing useless information.

(ii) *Feature Extraction*: Relevant features, in this case, words, are extracted from the pre-processed text. The most common way to do this is using techniques such as bag-of-words and n-grams.

(iii) *Model Development*: Train a machine learning model (e.g., Random Forest or Decision Tree) or a deep learning model to learn from data and later classify new unseen text sentiment correctly.

(iv) *Sentiment Classification Evaluation*: Use other models and adjust models to get the best performance possible.

Even though landscape images are included in “all” datasets and even though almost all authors used outdoor images in their work, those images showed a wide range of scenarios, many of which included man-made structures, people, or plants/animals close-ups. In the present work, we focused only on landscapes – natural outdoor scenes. We also propose ensemble/stacking modelling, which typically allows to increase in the accuracy of predictive

analytics and data mining applications. In this context, ensemble modelling or fusion is the act of executing two or more related but separate analytical models and then combining the results into a single score or spread. In the present case, we can relate different results from different/complementary models to achieve the best accuracy or to complement information. Examples of these techniques can be found in [18] or in [25].

The next section will briefly describe the sub-datasets created to test the model and the pre-processing steps applied to the data before analysing it.

2.3 DATASET

In the present study, we use three sub-datasets from two well-known datasets. The first sub-dataset (a) is going to be used to develop *Image Sentiment Classifier* models (see Section 2.4.1), being constructed as a restriction to the Flickr dataset (Flickr dataset available at: https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html) [26]. The second sub-dataset (b) was used to develop *Text Sentiment Classifier* models (see Section 2.4.2), being constructed from the T4SA dataset (T4SA is available at: http://www.t4sa.it/#dataset) [20]. Finally, the third sub-dataset (c) was used to build the integrated model, the *Multimodal Sentiment Classifier* model (see Section 2.4.3), being also formed using the T4SA dataset.

The decision to use these two well-known datasets (Flickr and T4SA) is based on the lack of a single dataset that offers the simultaneous ground truth for text and image sentiment. Further, it was used sub-datasets of those, as the full datasets have outdoor and indoor images, man-made scenarios, scenarios with persons etc., while we are only focused on landscapes. In more detail, the development of the *Image Sentiment Classifier* models involved the use of the Columbia Multilingual Visual Sentiment Ontology (MVSO) dataset, created from the Flickr dataset. This database covers a visual sentiment ontology consisting of 3,244 adjective-noun pairs and SentiBank, which is a set of 1,200 trained visual concept detectors, providing a mid-level representation of sentiment, associated training images acquired from Flickr, and a benchmark containing 603 photo tweets covering a diverse set of 21 topics [26], being available at https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html.

A subset was derived from Flickr dataset, specifically choosing landscape photos. This sub-dataset is designated as (a) ***FlickrColImg***. Since in the original dataset, the sentiment polarity of the images changes between -2 and 2, we considered for our models that a negative sentiment is between -2 and -0.5, a neutral sentiment is between -0.5 and 0.5, and a positive

sentiment is between 0.5 and 2. With 153k landscape images, the *FlickrColImg* sub-dataset is not balanced, so, ~80k balanced images were randomly selected from those 153k images (see Table 2.1).

The second and third dataset is based on the T4SA dataset [20], available at http://www.t4sa.it/#dataset. The dataset's collection process took place in 2016 and 2017, being the total number of tweets in the dataset of ~3.4 M, corresponding to ~4 M images, as each tweet may have more than one image. Each tweet (text and associated images) has been labelled according to the sentiment polarity of the text, namely negative (-1), neutral (0), or positive (1). The dataset's authors removed corrupted and near-duplicate images and selected a balanced subset of images, named B-T4SA. Figure 2.1 illustrates the architecture of the T4SA data collection process.

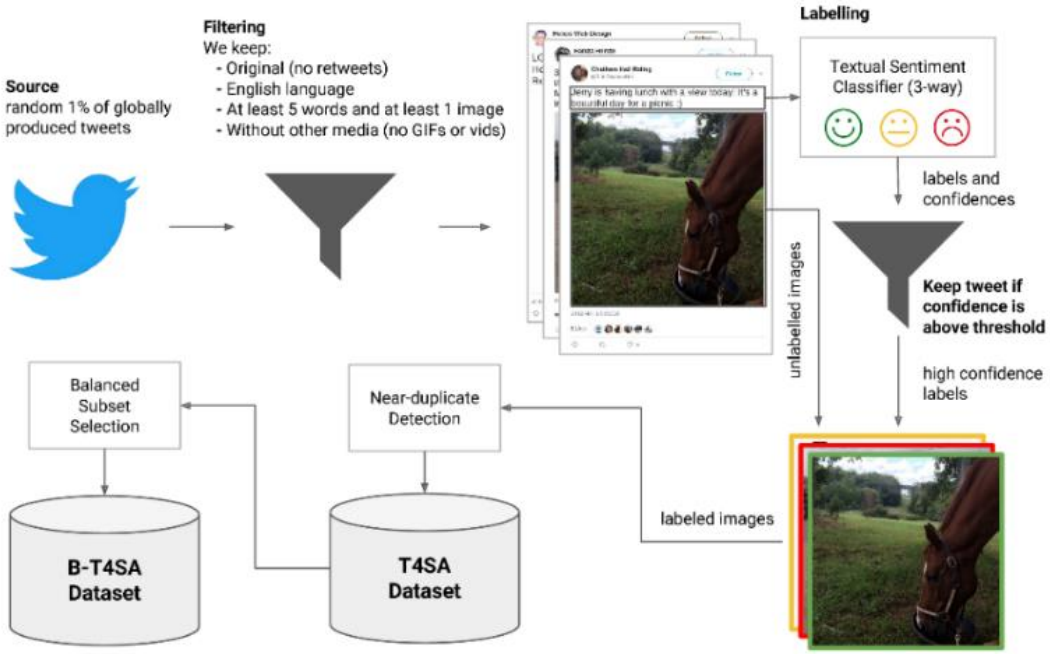

Figure 2.1. T4SA dataset collection process [20]. Image retrieved from http://www.t4sa.it/#dataset.

It is important to stress again that this dataset has outdoor and indoor images, where the outdoor images include landscapes, man-made images, images with persons etc. In our study, a sub-dataset of text posts was selected from the B-T4SA. These were used to train the *Text Sentiment Classifier* models, and it was designated as (b) ***B-T4SAtext***, consisting of 379k unbalanced samples, from which 50k balanced samples were randomly selected (see Table 2.1. Summary of the used data.). To this sub-dataset was applied the following pre-processing steps for text analysis: (i) the replacement of emojis/emoticons for words (e.g., 😍 was replaced by "smiling face with hearts"); (ii) convert all text to lowercase; (iii) remove stop

words (e.g., 'i', 'me', 'after', 'moreover'), proves useful as they do not contribute to sentiment analysis and their exclusion avoids unnecessary computations; (iv) removed HyperText Markup Language (HTML) tags, images, mentions, links, punctuation etc., because they do not carry sentiments; (v) apply a lemmatizer, which removes inflectional endings from a token to turn it into the base word lemma (e.g., the word 'dancing' would be lemmatized to 'dance'); and (vi) apply stemming, which is the process of removing suffices from words to obtain their root form (e.g., the word 'dancing' would be stemmed to 'danc'). Both stemming and lemmatizations serve the purpose of reducing word forms to their base or root forms to generalize the words, resulting in more accurate predictions in sentiment detection.

The final sub-dataset was designated (c) ***B-T4SAland***, consisting of only 850 samples (see Table 2.1. Summary of the used data.), being once again based on B-T4SA. For this dataset, the 850 posts (including image-text) were randomly selected, ensuring that the photos/images were of landscapes. Once the T4SA did not provide sentiment labelling (ground truth) for the images, those photos were presented and classified by a group of 6 persons (3 male, 3 female), with ages between 21 and 55 years, all with Portuguese nationality. The dataset was filtered to include only those cases where a minimum of five out of the six individuals unanimously agreed on the sentiment classification (positive, negative, or neutral). For each of those selected landscape images, the corresponding text was selected, and the respective sentiment label was retrieved from the B-T4SA dataset. This approach resulted in an unbalanced (sub-)dataset and other relevant occurrences, such as, the diversity of images/text sentiments pairing, where positive/neutral/negative classified images were paired with positive/negative/neutral texts classification. This is the reason why Table 2.1 shows, e.g., 639 images vs. 252 texts with positive sentiment or 96 images vs. 571 texts with neutral sentiment.

Table 2.1. Summary of the used data.

Sentiment	**Image -** *FlickrColImg*		***Text*** *- B-T4SAtext*		*B-T4SALand*	
	Original	**Balanced**	Original	**Balanced**	**Image**	**Text**
Positive (+)	89,746	**27,278**	127,086	**16,667**	639	252
Negative (−)	36,480	**27,278**	21,643	**16,667**	115	27
Neutral (=)	27,278	**27,278**	230,471	**16,667**	96	571
Total	153,504	**81,834**	**379,200**	**50,001**	**850**	**850**

To summarize, the number of images and text data divided per sentiment for each sub-dataset is represented in Table 2.1. Examples of photos are shown in Figure 2.2, upper row images from Flickr and bottom rows *B-T4SAland*, from left to right columns show positive, neutral, and negative sentiments. Examples of text are presented in Table 2.2, negative, positive, and neutral for the original post and the pre-processed text. The next section introduces the developed models.

Figure 2.2. Examples of landscape images used for training and testing where the upper row images are from Flickr and the bottom rows from B-T4SAland. From left to right are images classified as positive, neutral, and negative.

Table 2.2. Examples of texts used for the models' development.

Tweet/Sentiment	**Text**	**PreProcessed Text**
Negative (−)	"RT @Reuters: Special Report: Iraq militia massacre worse than U.S. acknowledged. https://t.co/KhzqwloPam https://t.co/8HtpZYQZQh"	special report iraq militia massacr wors u acknowledg
Positive (+)	"Have a look at the most spectacular #NationalParks in #California. #traveler https://t.co/wiIfw7nnH7 https://t.co/sxnJnQH0cC"	look spectacular nationalpark california travel
Neutral (=)	"Live Next Door To Parents https://t.co/FfJrf5n8D1 https://t.co/2bZRdnkPQY"	live next door parent

2.4 LANDSCAPE SENTIMENT CLASSIFICATION MODEL

The *Landscape Sentiment Classification model* (LSCm) is used to extract sentiments from photos-images and texts, following the principles presented in [19]. However, our model exhibits distinctions when ensembles are employed. Namely, image and text sentiment classification results from the combining of various methods, i.e., components are combined into an ensemble to yield the ultimate result. In this context, the possible outputs are (see

Figure 2.3. Proposed model for LSCm.): (i) sentiment resulting from the image (*isc*), generated by the *Image Sentiment Classification* (ISC) block; (ii) sentiment resulting from the text (*tsc*), generated by the *Text Sentiment Classification* (TSC) block; (iii) sentiment resulting from the combination of image and text (*msc*), generated by the *Multimodal Sentiment Classifier* (MSC) block; and (iv) sentiment discrepancy between image and text (*dis*). So, for each paired (image - text) the model returns the following vector [*isc*, *tsc*, *msc*, *dis.*].

Figure 2.3 shows a simplified diagram block of the model. At the top is presented the generic LSCm, where "+" represents a positive sentiment, "-" a negative sentiment, "=" the neutral sentiment, and *dis.* is the discrepancy between the sentiment returned from the image and the text, for each pair of image-text presented in the input. In the middle of the figure are the image and text sentiment classifier blocks, ISC and TSC, respectively. The bottom part of the figure represents the schematized combination of all the inputs to return the final output in the Multimodal Sentiment Classifier (MSC) block. Details of these blocks will be presented in the following sections.

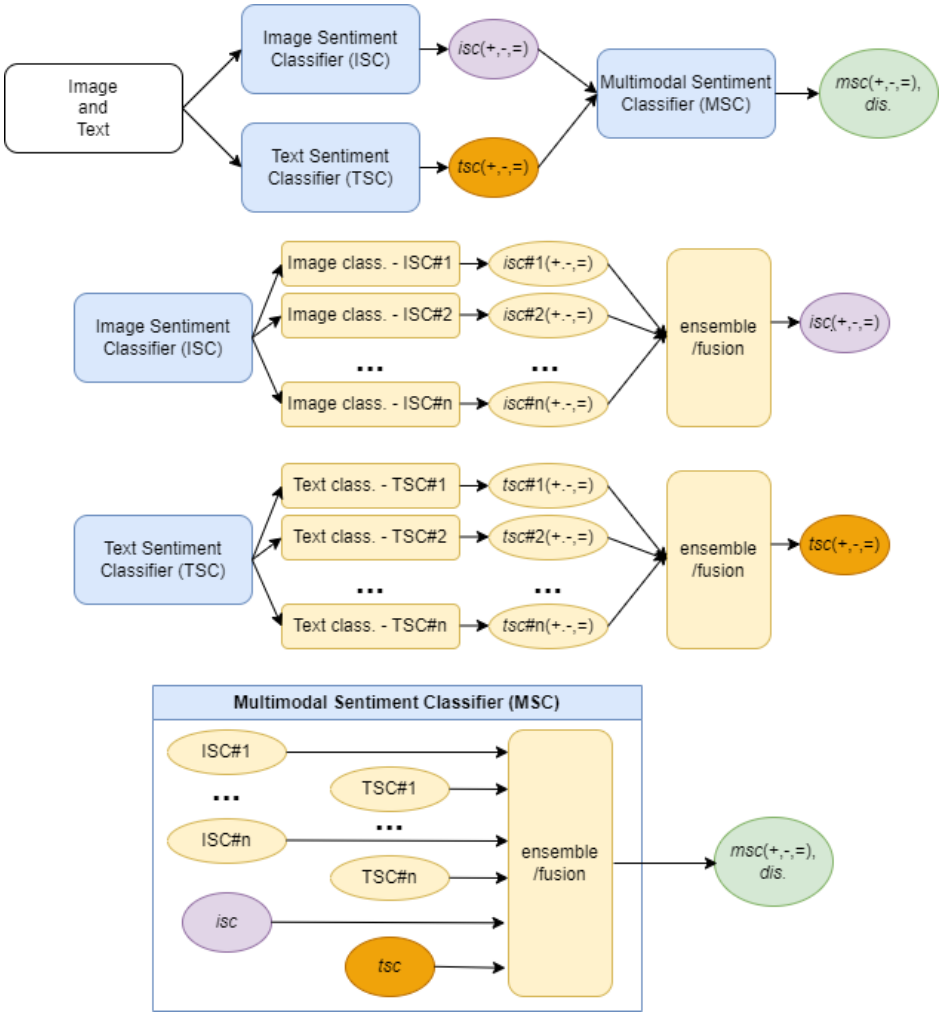

Figure 2.3. Proposed model for LSCm.

2.4.1 IMAGE SENTIMENT CLASSIFIER BLOCK

The landscape Image Sentiment Classification block is based on the work of Oliveira et al. [21], which creates a deep learning (DL) sentiment classifier model based on a traditional DL architecture backbone (such as Visual Geometry Group (VGG), ResNet, Inception etc.) followed by a "handcrafted" network head, based on fully connected layers.

In this scenario, three primary models are introduced (predefined), distinguished by the applied backbone architecture. Specifically, one utilizes the DenseNet121 (121 layers) architecture as the backbone [26], another uses the Xception architecture (71 layers) [27], and the third one uses the ResNet-50 architecture (50 Layers) [28]. It is worth noting that all the backbones were trained using the ImageNet dataset. Different strategies of fine-tuning were used for these models, including different numbers of fully connected layers at the network's head.

The models are depicted in the 2nd row of Figure 2.3 and Table 2.3 shows the optimized hyperparameter values, each pre-defined architectures sub-model summarized as follows:

(a) **ISC_DL#1:** the backbone is a DenseNet121. Furthermore, let $L_{i,j}$ represent a dense layer, where *i* is the layer number and *j* is the number of units. Then, in ISC_DL#1, the head has 5 layers. In the initial layer, L_{in}, the number of units equals the number of outputs of the backbone. The second dense layer has *n* units with X_2 activation function ($L_{2,n}$) and dropout D_{d2}, a dense layer $L_{3,m}$ with D_{d3}, and a dense layer with 24 units ($L_{4,24}$), all with X_i activation functions. The last layer has 3 units with a softmax activation function (Ls). The layer of 24 units is based on our hypothesis derived from Plutchik's wheel of emotions. We hypothesize that due to the importance of color in sentiment analysis and the relation between emotion and sentiment, and once there are 24 emotions in Plutchik's wheel, it is expected that a layer of 24 units can help the network to learn those emotions and relate them with the 3 sentiments, which appears in the last layer. I.e., it will allow to classify the sentiment of the image into positive, negative, or neutral according to the responses of these "emotions".

(b) **ISC_DL#2:** the backbone is Xception and the head has 6 layers. The first dense layer has L_{in}, the number of units equals the number of outputs of the backbone. Then, a dense layer with *n* units ($L_{2,n}$) and D_{d2}. The third layer has also *m* units and D_{d3}. The fourth layer has *o* units, with D_{d4}, then the dense layer with 24 units ($L_{5,24}$), all layers with X_i activation functions, the last layer is the Ls.

(c) **ISC_DL#3:** the backbone is ResNet-50 and the head has 5 layers. The first dense layer is L_{in}, the second layer $L_{2,n}$ with D_{d2}, the third layer $L_{3,n}$ with D_{d3}, a fourth layer $L_{4,24}$, all with X_i activation functions, and the last layer is the Ls.

These architectures were fine-tuned (see Table 2.3) and several hyperparameters were tested, such as number of units, drop-out rate, batch size, number of epochs etc. (see section 2.5). The only fixed values were the penultimate (with 24 units) and the last (with 3 units) layers, and the softmax activation function.

Them, the results from the three classifiers (ISC_DL#1, ISC_DL#2, and ISC_DL#3) are to be combined/ensembled to obtain a final classification. So, the inputs of the ensemble sub-block will be the predictions made by the individual models, resulting from the softmax function, and the output will be the final prediction, *isc*. Ensembling leverages the idea that combining the strengths of multiple models can result in improved overall performance and accuracy, compared to using a single model, as well as a better generalization (reacts better to unseen data than single models), reduces "overfitting", and increases the robustness of the results obtained. For the aggregation of sentiment classification (ensemble sub-block), it was used:

(i) ***Random Forest (ISC_RFa)*:** *k* estimators (see Table 2.3), Gini impurity as criteria function to measure the quality of split, and the minimum number of samples required to split an internal node was 2.

(ii) ***Neural Network (ISC_NNa)*:** three dense layers, where the first layer has 9 units (L_{in}), then a dense layer with *n* units ($L_{2,n}$) with X_i activation functions, and the third layer the Ls.

A search tuner function was used to find the best (hyper-)parameters for the proposed models (see Table 2.3).

2.4.2 TEXT SENTIMENT CLASSIFIER BLOCK

In the landscape *Text Sentiment Classification* block, the first step consists of converting the text (see Section 2.3) into structured data, in a way that can be used by machine learning (ML) methods. To achieve this, a Bag of Words (BOW) was applied to extract 5,000 features, with one to two n-grams, from the processed texts. Figure 2.4 illustrates the feature extraction from text for sentiment classification.

Using a fine-tuning procedure, the BOW was used to train two ML models (see Table 2.3):

(a) ***Random Forest (TSC_RFc)***: created with *k* estimators and Gini impurity as criteria function to measure the quality of the splits**.**

(b) ***Neural Network (TSC_NNc)***: 4 dense layers where the first has 5,000 units (L_{in}), then two layers with *n* units ($L_{2,n}$) units and *m* units ($L_{3,m}$) are added, with X_i activation functions. The last layer uses 3 units with a softmax activation function (Ls), used to predict the probability of a text carrying a positive, negative, or neutral sentiment.

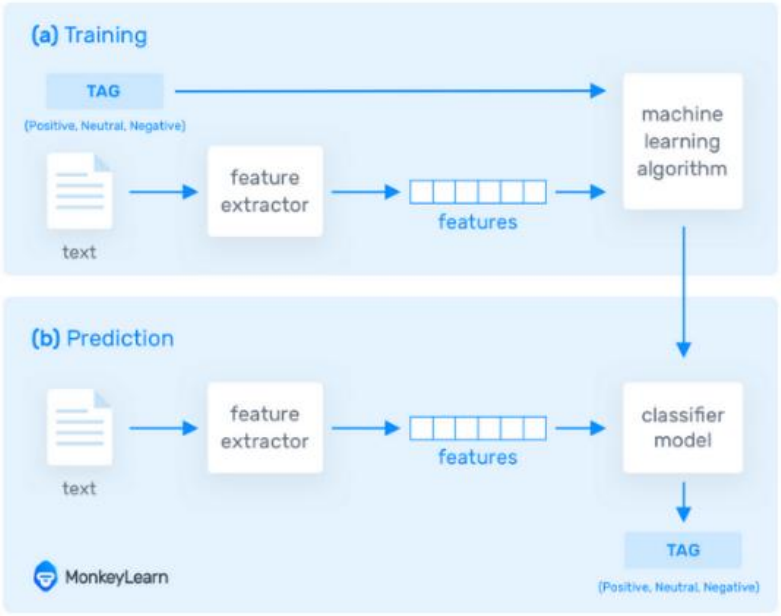

Figure 2.4. Features extraction from text for sentiment detection. Image retrieved from http://www.datasciencelovers.com/tag/tf-idf/.

Furthermore, the (c) Natural Language Toolkit (NLTK) method, available in the literature at https://www.nltk.org/, was used as a third model (TSC_NLTK).

The block diagram of the proposed TSC block is presented in the 3rd row of Figure 2.3. For the aggregation of text sentiment classification (the ensemble sub-block) was used:

(i) ***Random Forest (TSC_RFt)*****:** *k* estimators and "Gini criteria" function;

(ii) ***Neural Network (TSC_NNt)*****:** 4 dense layers, the first layer has 9 units (L_{in}), then a dense layer with 3 units ($L_{2,n}$), a third layer with *m* units ($L_{3,m}$) all with X_i activation function, and the fourth layer the Ls.

As before, a search tuner function was used to find the best (hyper-)parameters for the proposed models (see Table 2.3).

2.4.3 MULTIMODAL SENTIMENT CLASSIFIER BLOCK

The *Multimodal Sentiment Classifier* block, as largely mentioned in the text, is based on classifications resulting from text and image. Going back to the block diagram in Figure 2.3, bottom row, the ensemble block has 3 inputs from ISC and 3 inputs from TSC. Also, *isc* and *tsc* are used to compute the *discrepancy*:

(a) ***If the input sample (image and text) has similar sentiments***, i.e., *isc* = *tsc*, then *msc* = *isc* = *tsc*, and LSCm = [*isc*, *isc*, *isc*, *dis.* = *0%*] is the resulting output.

(b) ***If the input sample (image and text) has different sentiments*** (*isc* ≠ *tsc*), considering μ_I the average results (between 0-100) from all ISC models that returned the same sentiment, μ_T the average of all TSC models that returned the same sentiment, and the threshold μ_t = 85 (empirically calculated):

(b.1) If $\mu_I \geq \mu_t$ and $\mu_T \geq \mu_t$ then both ISC and TSC are certain of the sentiment and, in this case, most probably the person who posted the image-text intended different sentiments. The ensemble block is not computed, resulting LSCm = [*isc*, *tsc*, ×, *dis.* = 100%]. This case means that ***the image and text have different sentiments.***

(b.2) If $\mu_I < \mu_t$ and $\mu_T < \mu_t$ and $\mu_I \sim \mu_T$ (~ means approximately, with a difference of ± 2 percentual points) then image and text bring the same contribution, i.e., *dis.* = 50%. In this case, the ***ensemble is computed,*** being LSCm = [*isc*, *tsc*, *msc*, *dis.* = 50%]. This case means that ***the image and text are indeterminate if they have the same sentiments.***

(b.3) All other situations mean that ***ensemble must be computed***, once there is no certainty about what the person intended to post. Resulting in this case the *dis.* = (μ_I + μ_T)/2, and LSCm = [*isc*, *tsc*, *msc*, *dis.* = ((μ_I + μ_T)/2) %]. This case means that ***the image and text complement the final result***.

As there is no balance in the sentiment classes, the Stratified K-Fold cross-validation technique was used to train and evaluate the MSC model. The ensemble block is computed following the same principles presented before, namely:

(i) ***Random Forest (MSC_RFe)*:** *k* estimators and Gini as criteria function;

(ii) ***Neural Network (MSC_NNe)*:** 4 dense layers, where the first layer has 18 units (L_{in}), a dense layer with *n* units ($L_{2,n}$), a third layer with *m* units ($L_{3,m}$), with X_i activation function, and the fourth layer the Ls with softmax as activation function.

A search tuner function was used to find the best structure and parameters, see Table 2.3. In the following section test and results will be presented.

2.5 TESTS AND RESULTS

The procedure implemented to evaluate the model is divided into three steps: (a) evaluate image sentiment classification, (b) evaluate the text sentiment classification, and (c) evaluate the ensemble of both text and image models, i.e., evaluate the LSCm model. The models were trained and tested using a combination of the Google Colab (with 12.7GB de RAM (Random Access Memory) and 107.7GB disk space) and Kaggle platform (with 14.8 GB RAM e 2x GPU (Graphics Processing Unit) NVIDIA T4). For the Random Forest models, the Gini impurity was the criteria function to measure the quality of split and the minimum number of samples required to split an internal node was 2. The model was trained in two ways: (a) by using the sentiment values of -1 (negative), 0 (neutral), and 1 (positive), predicted by the individual models, and (b) by using the sentiment-predicted probabilities from the individual models. As the results were very similar, so the option (b) was chosen.

Table 2.3 shows the ISC, TSC and MSC parameters, hyperparameters, and accuracies of the models. In more detail, the first column shows the used models, the second column summarizes the number of units used in the layers (resulting from the optimizations), as well as the estimators, and the third column shows how many layers the backbone has, as well as some hyperparameters used to train the new data, dropout, and activation functions. The accuracy is presented in the last column.

Lines 3 to 6 present the information related to the ISC models, lines 8 to 11 the information related to TSC models, line 13 the results of the MSC model, and line 14 the LSCm's results. The last line shows the results from the analysed image-text, in percentage, the different types of combinations of sentiment classification, where: *dis.* = 0% corresponds to ISC and TSC reporting the same sentiment; *dis.* =100% corresponds to ISC and TSC reporting different sentiments; and other *dis.* values when the sentiment is undetermined, i.e., the image and text point to opposite sentiments, but this is not completely clear.

Going back to the evaluation procedure of the model, in more detail: (a) the image models were trained with *FlickrColImg*, composed of 81,834 images (see Table 2.1), where 80% of

those images were used for training and 20% for testing. As the dataset had images with different scales, (a.1) image pixel values were normalized between 0 and 1 and (a.2) resized to 224 × 224 pixels. The (b) text models were trained with *B-T4SAtext*, composed of 50,001 text tweets, where 80% of the samples were used for training and 20% for testing. The (c) MSC was trained using *B-T4SALand* and a stratified K-fold cross-validation technique, but only using the samples where the image and the text have the same sentiment, resulting in 273 samples. For the final results, LSCm's results (see Table 2.3, 14th line), the inference is done over the already trained models, mentioned above, using *B-T4SALand* dataset, with 850 samples.

Table 2.3. ISC, TSC, and MSC parameters, hyperparameters, and accuracy of the models.

Model	***Number units / estimators***	***Backbone layers trained // hyperparameters + dropout + activation function***			***Accuracy***
Image Sentiment Classification					
ISC#1 (DL)	*n = 2048* *m = 1024*	8 Opt: SGD (1e-2) Epochs: 20 Batch size: 32	d_1=50% d_2=50%	X_i = ReLU *i*={1,…,4}	54.55%
ISC#2 (DL)	*n = 4096* *m = 4096* *o = 2048*	8 Opt: SGD (1e-2) Epochs: 20 Batch size: 32	d_1=70% d_2=50% d_3=30%	X_i = ReLU *i*={1,…,5}	54.79%
ISC#3 (DL)	*n = 1024* *m = 512*	All Opt: SGD (1e-4) Epochs: 20 Batch size: 32	d_1=50% d_2=50%	X_i = ReLU *i*={1,…,4}	54.73%
ISC – RFa **ISC** – NNa	*k = 100* (est.) *n = 57*	Opt: Adam (7e-4) Epochs: 20 Batch size: 32	—	X_i = ReLU *i*={1,…, 3}	55.46% **55.89%**
Text Sentiment Classification					
TSC#1 (RF)	*k = 100* (est.)	—	—	—	90.10%
TSC#2 (NN)	*n = 300* *m = 100*	Opt: SGD (1e-2) Epochs: 10 Batch size: 8	—	X_i = ReLU *i*={1,…, 3}	88.55%
SC#3 (NLTK)	—	—	—	—	84.34%
TSC – RFt **TSC** – NNt	*k = 100* (est.) *n = 100* *m = 20*	Opt: SGD (1e-2) Epochs: 10 Batch size: 2	—	X_i = ReLU *i*={1,.., 3}	91, 95% **92,10%**
Multimodal Sentiment Classification					
MSC - RFe **MSC** – NNe	*k = 100* (est.) *n = 100* *m = 24*	Opt: Adam (1e-2) Epochs: 10 Batch size: 32	—	X_i = ReLU *i*={1,…, 3}	**100.00%** 98.46%
LSCm	—	—	—	—	**78.75%**

Sentiment Disparity Between Image and Text

	Ground-Truth	Prediction
dis.=0%	32,11%	38,00%
dis.=100%	4,47%	4,24%
dis.= #%	64,41%	57,76%

In Table 2.3, the high accuracy of the MSC models is justified by the fact that they are solely employed to determine and enhance sentiment accuracy based on the results of the six individual models (3 TSC and 3 ISC). So, for nearly 100% of the cases, the previously predicted sentiment will remain the same. At this point, it is also important to stress again how the accuracy of LSCm is computed, as it aggregates results from the MSC model (a), (b.2),

and (b.3), and checks these predictions against the ground truth, once (b.1) image and text reflect or could reflect different sentiments.

One of the main focuses of the chapter is to understand if the image and text in a post represent the same sentiment, and present the alignment between real sentiments (ground truth) and predicted sentiments, for both text and image data. In the figure, "-1" are negative sentiments, "0" are neutral sentiments, and "1" are positive sentiments. At the top of Figure 2.5, the confusion matrices compare the real *vs.* predicted sentiments of the text (left) and image (right) samples, while at the bottom the confusion matrices show the difference in real and predicted sentiment between the text and image.

By analysing text and images isolated (Figure 2.5, top), it is evident that the main error of the model is to assign positive images to neutral images. Furthermore, by observing the bottom row, where we can compare our predictions and the real results for the text-image combination, we can see that the results are quite similar. One example is "Image *vs.* Text - True" matrix, which reflects that 192 pairs show the same positive sentiments, 438 show indeterminacy (i.e., both image and text can report the same sentiment or opposite sentiments), and 9 show different sentiments. Now looking at the same line but in the confusion matrix for "Image vs. Text - Predictions", we have respectively 195, 388 and 20, being only 3 more pairs are assigned to positive sentiment.

If we analyse the discrepancy between the ground truth data and our previsions, Table 2.3 bottom line, we see the same trend. So, results that point to an indeterminacy of the image and text sentiment should be used to complement the global sentiment of the post. The left side of Figure 2.6 shows the discrepancy bar chart (3 sentiments) and on the right the prediction confidence percentage level bar chart (3 sentiments).

Table 2.4 exemplifies the *confidence* in how the MSC predicts a sentiment. The MSC model uses the 6 individual estimators to determine a certainty probability for the final sentiment so that, if all of those return the same predictions then the final sentiment and the MSC is the same (normal behavior), which results in a high accuracy of the model. In the example, the model predicts the final sentiment with 90% certainty of being positive.

We also tested the model for 2 sentiments, i.e., from the sub-datasets *FlickrColImg*, *B-T4SAtext*, and *B-T4SALand* we removed the neutral samples and retrained the models again, using the same parameters, except for the output layer that was adapted in accordance. The results achieved show the same tendency as with 3 sentiments, as can be observed in Figure 2.7 and Figure 2.8.

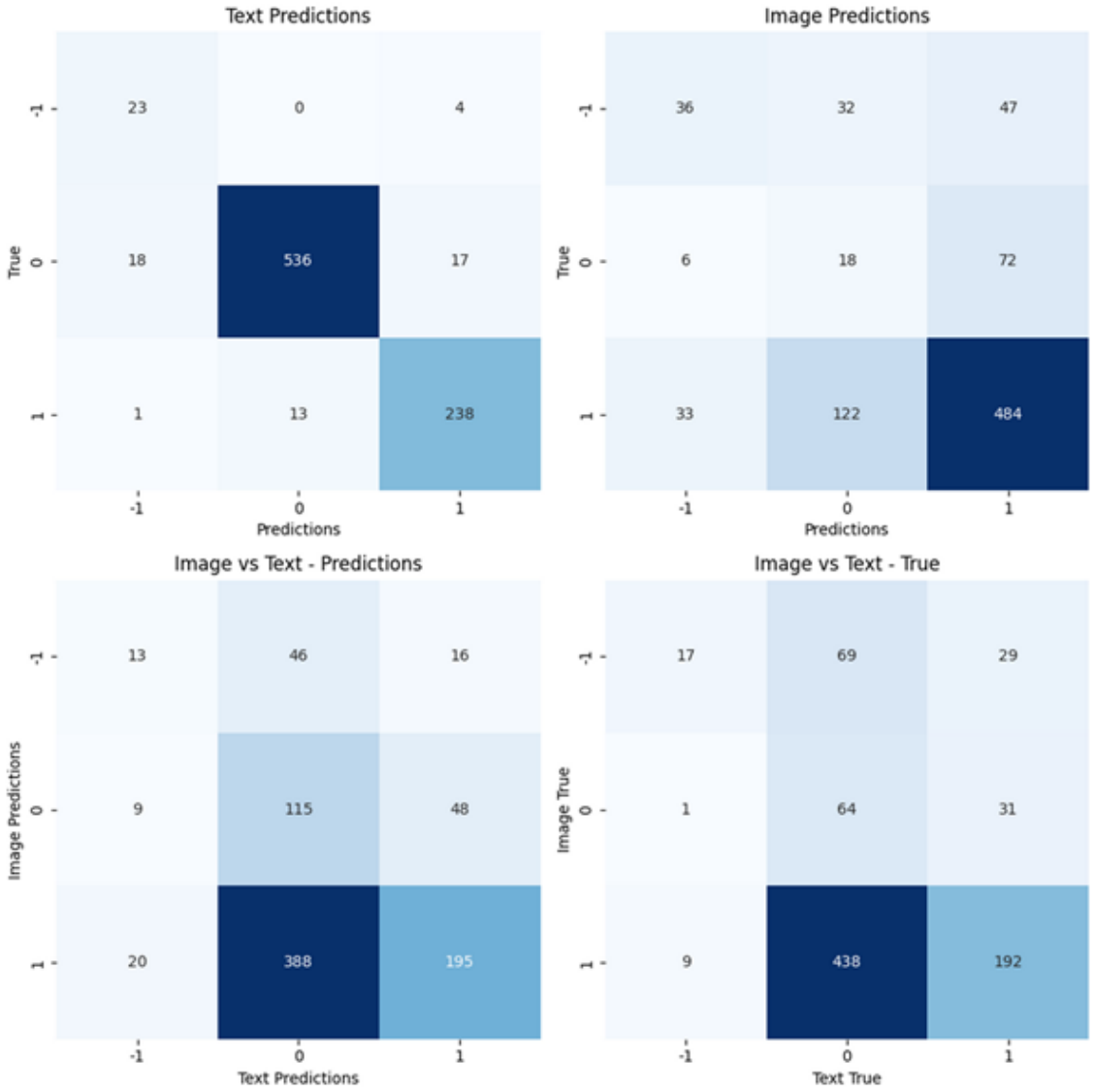

Figure 2.5. Confusion matrices of images, text, and text-image classifications (3 sentiments).

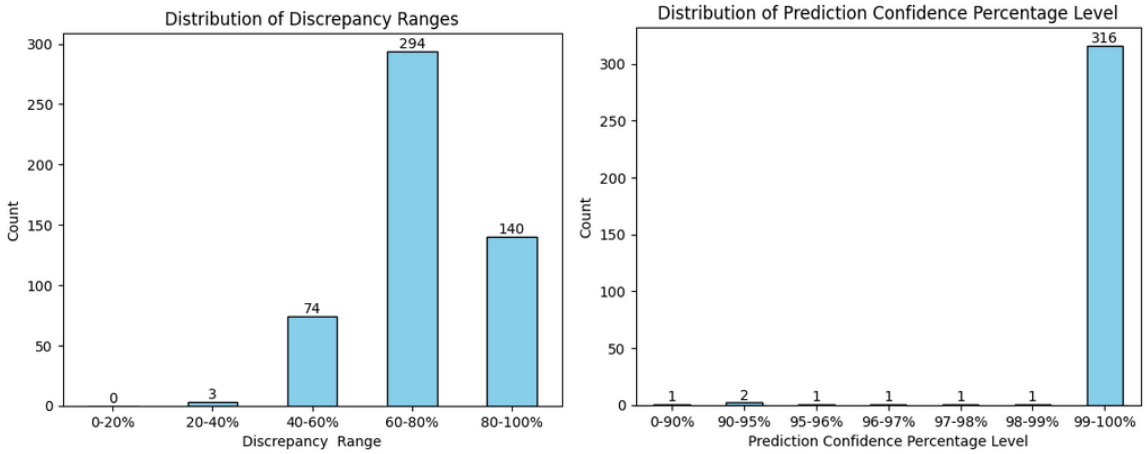

Figure 2.6. On the left, discrepancy bar chart (3 sentiments) and, on the right, the prediction confidence percentage level bar chart (3 sentiments).

Table 2.4. Example of prediction confidence percentage level sample.

	NEGATIVE	**NEUTRAL**	**POSITIVE**
ISC#1	0%	25%	75%
ISC#2	10%	10%	80%
ISC#3	10%	40%	50%
ISC (ensemble)	0%	18%	82%
TSC#1	10%	20%	70%
TSC#2	25%	25%	50%
TSC#3	0%	70%	30%
TSC (ensemble)	0%	26%	74%
TSC - ISC	POSITIVE – POSITIVE		
MSC	0%	10%	90%
LSCm	POSITIVE		

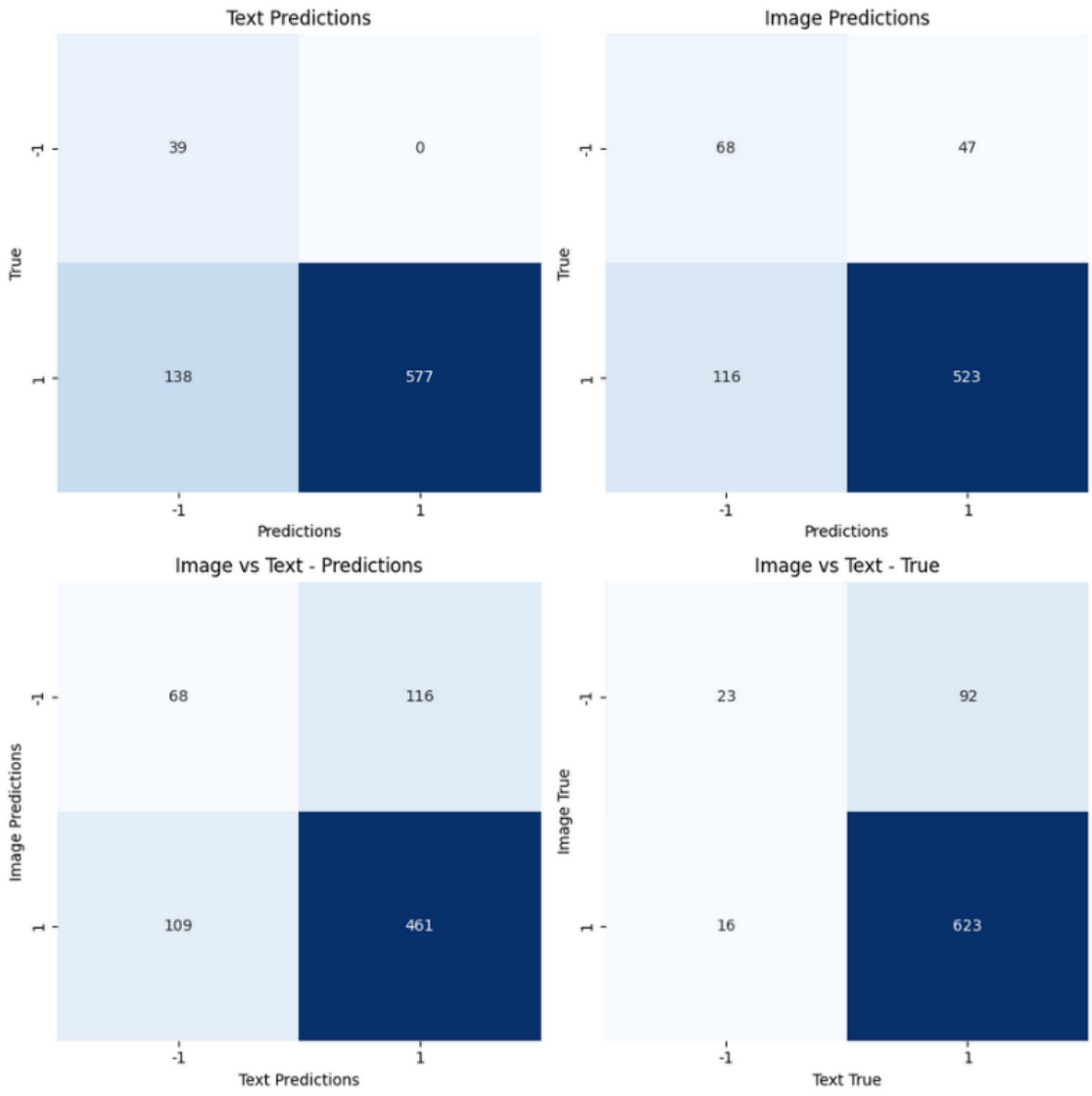

Figure 2.7. Confusion matrices of images, text, and text-image classifications (2 sentiments).

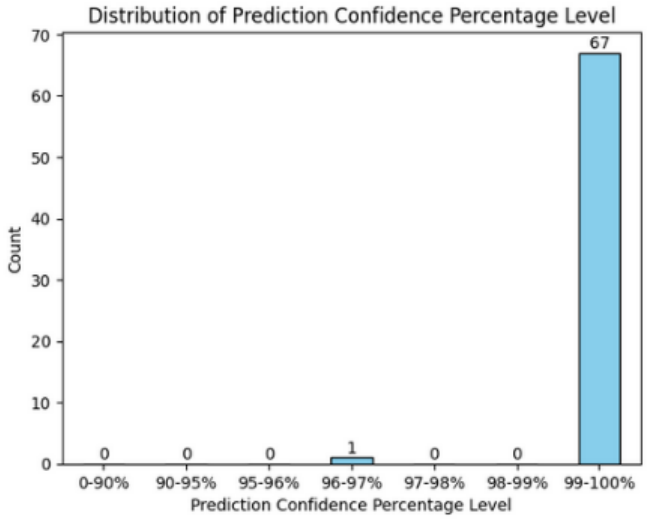

Figure 2.8. Prediction confidence percentage level bar chart (2 sentiments).

2.6 CONCLUSIONS AND FUTURE WORK

This work presented initial research on sentiment classification for text and image data, especially in landscape-related social media posts. Our study and analysis demonstrated the effectiveness of the proposed text and image sentiment model, highlighting its potential applications and implications. Our results indicate that the model achieves high accuracy in sentiment classification from image-text data. The pre-processing techniques, incorporation of deep learning models, and advanced feature extraction techniques used have led the system to get this high accuracy. Furthermore, the experiments on image sentiment detection have shown promising results, demonstrating that the system can classify sentiments from landscape images.

The ensemble blocks allow the system to leverage the complementary information provided by the textual and visual models, leading to enhanced sentiment classification performance. This is very important when a multiple model approach is used, in this case in sentiment analysis tasks.

In summary, this initial prototype contributed to the understanding of sentiment detection from pair image-text data, with a focus on landscape images. Although the accuracy of the system is still not optimal the potential of the model was shown. Continuing to improve and refine landmark sentiment classification, can open up new possibilities for enabling more sophisticated and nuanced sentiment analysis in interfaces and/or robots.

Looking ahead, there are several directions for future research. The focus should be on improving the sentiment detection model in images and getting more and, also extremely relevant, better datasets of image-text related to landscapes. We also intend to test the same models in pairs text-image of indoor, and outdoor man-made scenes as well as scenes with persons and finally combine all the information.

3

MULTIMODAL SENTIMENT CLASSIFIER FRAMEWORK FOR DIFFERENT SCENE CONTEXTS

3.1 INTRODUTION

The goal of Human-Centered Artificial Intelligence (HCAI) is to create technologies that assist people in carrying out various daily tasks, while also advancing human values, such as rights, fairness, and dignity [5]. As an interdisciplinary area involving computer science, psychology, and neuroscience, HCAI aims to achieve a balance between human control and (complete) automation by improving people's autonomy, well-being, and influence over future technologies. Affective Computing (AffC) is a related field that combines the fields of sentiment analysis and emotion recognition. AffC is backed by a variety of physical information types, including text, audio (speech), visual data (such as body posture, facial expression, or environment), and physiological signals (such as electroencephalography or electrocardiograms). AffC may be developed using either unimodal or multimodal data within this framework [6].

HCAI and AffC have a wide range of applications. For example, in *lato sensu*, a machine needs to be designed to cooperate with or learn how to function in interpersonal relationships with people. Emotions and sentiments are fundamental to human-machine relationships, and any robot communicating with humans must incorporate them. Conversely, social media platforms are becoming increasingly significant in today's digital marketing landscape, as they influence individuals to travel, look for and purchase goods, alter their lives, and alter their perspectives on many topics. The sheer number of daily posts on various platforms has made

it necessary and simultaneously possible to monitor, evaluate, and comprehend the mood, sentiments and emotions such messages convey.

Considerable progress has been achieved in the area of sentiment analysis [29, 30, 31, 32], including, but not limited to, text, image, and text-image posts (even in cases when the image and words together may convey a different meaning), music, and video analysis, or even for robots, that can read facial emotions and improve human-human relations.

The quickest and easiest kind of publications to get people to click, buy, and read about a specific topic or product include short texts, images, and/or videos. This explains why social media platforms like Instagram and X (formerly known as Twitter) have gained much traction in recent years.

There are two main categories of posts where the pair text-image is connected as follows:

(i) *where the text is (clearly) complemented by an image(s) of a person or group of persons*. In those cases, the environment (scene) is the background and the person(s) are in the foreground, i.e., they are looking in a (semi-)frontal manner at the camera, so their facial expressions or body posture have a powerful influence on the sentiment carried [7].

(ii) *where the text may or may not be complemented by the image(s)*. In those cases, there are no persons in the scene, or if existing they are the "background", and the scene is the "foreground". In those cases, color, texture, edges, line type, orientation etc. are important features in the attraction and sentiment that the image carries [8].

This chapter focuses on the second case, i.e., detecting sentiments in posts relating to pair text-image(s) without persons, or existing persons who are only in the "background". In this case, there isn't in the literature a good sentiment classifier for this type of image and text combination (see next section). To the best of our knowledge, sentiment classifiers that deal with this type of image use a holistic approach, i.e., the models are trained with all types of images, without considering their specific characteristics. Here the images are segmented into four categories, namely, "non-man-made outdoors" (ONMM), "man-made outdoors" (OMM), "indoor" (IND), and "indoor/outdoor with persons only in the background" (IOwPB). Each class has partial results that are combined, resulting in a final model, which replaces the holistic approach.

The present proposal for the *image sentiment classifier* (ISC) has been developed based on three deep-learning (DL) models, which were fine-tuned from one developed for the class ONMM. This work was presented in the previous chapter. The principle of using this category

to develop the baseline DL models was that it is a more generic category, it is expected to return a good baseline for the remaining categories. If dedicated DL models are developed for each category, the final ISC model will achieve higher accuracy. Nevertheless, the authors' investigation hypothesis is that when using 4 models/categories to develop the ISC, even if those are not fine-tuned for each category, the final result (accuracy) will be better than using a single holistic model.

Considering the above principle, it was applied an ensemble of the DL models to achieve the final ISC_ONMM model, ISC_ONMM = E{DL#1, DL#2, DL#3}, with E denoting the ensemble. The same principle and models, without additional tuning, were trained and applied to the other three categories (OMM, IND, and IOwPB). The ensemble of the models, ISC = E{ISC_ONMM, ISC_OMM, ISC_IND, ISC_IOwPB}, was then combined with machine learning (ML) models for the *text sentiment classification* (TSC), see [6] for more details, returning the final *multimodal sentiment classification* (MSC) model.

In this context, where the text and image(s) may or may not convey the same sentiment, the information about the sentiment discrepancy between text and image depends on the individual results (for each modality). This discrepancy is used to decide if the image(s) should complement the sentiment information or simply state that the text and image represent completely different sentiments. In those cases, probably, the intention of the user/poster is just to put an image to illustrate the post, not to reflect his/her sentiment, or to be sarcastic.

In summary, this work presents a multimodal text-image sentiment classifier framework. In the case of the image, the *Image Sentiment Classifier* has several classifiers trained with different segments (4), that can be compared with a *Holistic Image Sentiment Classifier* (HISC), trained with all images available. The sentiment results from the images are combined or not (depending on the discrepancy) with the text sentiment results, returning the sentiment classification and discrepancy between the image and text predicted sentiments.

The main contributions of this work include threefold: (i) the (single hyperparameter) image classification sentiment model that works in different scenes/environments (ONMM, OMM, IND, and IOwPB); (ii) the consideration of the discrepancy between the image and text sentiment, which is not present in the literature, and can be used to decide if the text or image should be used to complement the sentiment information or not (e.g., this particularly relevant for a post that may be sarcastic); and (iii) the framework that combines image and text sentiment classification, returning the multimodal sentiment classification attach with the discrepancy (text-image sentiment) metric.

The present section introduces the work's goal, section 3.2 presents the contextualization and state of the art, section 3.3 introduces the datasets used, section 3.4 details the models, and section 3.5 outlines the tests, results, and respective discussion. Finally, section 3.6 presents the conclusions and future work.

3.2 CONTEXTUALIZATION AND STATE-OF-THE-ART

The affiliation, warmth, friendliness, and dominance between two or more individuals are displayed when relationships are made, returned, or deepened [9]. Opposite, impersonal human-machine interactions impede more extensive communication and complicate the establishment of intimate or reciprocal relationships between people and machines, devices, or interfaces. Within this framework, automated data evaluation systems to determine the conveyed emotion are known as automatic emotion analysis approaches [10] (categorized, e.g., as happiness, sadness, fear, surprise, disgust, anger, or neutral) or sentiment [11, 12] (typically limited to positive, negative, and neutral).

Considerable advancements have been achieved in the field of sentiment categorization in recent years. Though sentiments and emotions are distinct entities, it should be noticed that they are interconnected [12, 13], i.e., emotions affect sentiments and sentiments affect emotions. The sentiment may be impacted by a variety of elements including attitudes, opinions, emotions, prior experiences, cultural background, personal views, age, or even gender. It is a mental attitude that is connected to a good, negative, or neutral evaluation or thinking about anything [13].

As an example, color can be considered an important feature for sentiment classification [8, 14, 15, 16]. Typically, images of beaches or oceans predominantly feature blue tones, evoking a sense of calm. In contrast, an image of a forest will highlight green tones, which are associated with harmony. Nevertheless, the meaning of a color can change depending on the context in which it is used; for example, the red color can sometimes represent anger, love, or frustration. Also, different individuals may perceive color differently on an emotional level, just like they do with music.

Different authors categorize emotions in a variety of ways and break them down into levels and sublevels [13]. Six fundamental emotions, which are typically accompanied by the neutral emotion, is the classification used by the renowned psychologist Paul Ekman [10]. This group of emotions is commonly used for facial emotion classification. Other authors have also put

forward different categories. For example, based on biological mechanisms, Robert Plutchik [17] defined eight basic/primary emotions: joy, trust, fear, surprise, sorrow, disgust, anger, and anticipation. Plutchik created a color wheel, known as Plutchik's wheel, to symbolize feelings, with a particular color assigned to each. Emotions in this instance may be categorized according to their intensities and combinations; that is, primary emotions are coupled in various ways to generate secondary and tertiary emotions, which are symbolized by various color tones and hues, for a total of 24 emotions.

In summary, Plutchik's wheel categorizes emotions into two primary sentiments: positive and negative. Joy, trust, anticipation, and surprise are examples of good sentiment; on the other hand, sorrow, contempt, fear, and rage are examples of negative sentiment. As it was discussed before [16], the division of emotions into positive and negative sentiments can be subjective and based on individual and cultural variables. Compared to emotion, which can shift quickly in reaction to shifting circumstances and stimuli, sentiment is often longer and more consistent [18].

Ortis et al. [13] provided a summary of sentiment analysis in images. The authors outlined the prospects and difficulties in the field and discussed the main problems. To classify the content of composite comments on social media, a multimodal sentiment analysis (text and image) model was published by Gaspar and Alexandre [19]. The three primary components of the technique are an image classifier, a text analyser, and a method that examines an image's class content, determining the likelihood that it falls into one of the potential classes. By combining deep features with semantic information obtained from the scene characteristics, the authors also assess how classification and cross-dataset generalization performance might be enhanced. The authors used the T4SA dataset [20], as the source of their study, which consists of three million tweets – text and photos – divided into three sentiment categories (positive, negative, and neutral), was the source of their study.

In [8], a color cross-correlation neural network for sentiment analysis of images was introduced. The architecture considers the relationships between contents and colors in addition to utilizing them concurrently. The authors collected color features from several color spaces using a pre-trained convolutional neural network to extract content characteristics and color moments. Then, using a sequence convolution and attention mechanism, they present a cross-correlation method to model the relationships between content and color features. This method integrates these two types of information for improved outcomes by enhancing the sentiment that content and color express.

In [21] the authors suggest a system to categorize the tone of outdoor photos that people post on social media. They examine the differences in performance between the most advanced ConvNet topologies and one created especially for sentiment analysis. The authors also assess how classification and cross-dataset generalization performance might be enhanced by combining deep features with semantic information obtained from the scene characteristics. Finally, they note that the accuracy of all the ConvNet designs under study is enhanced by the integration of knowledge about semantic characteristics.

A deep-learning architecture for sentiment analysis on 2D photos of indoor and outdoor environments was presented by Chatzistavros et al. [22]. The emotion derived from catastrophe photographs on social media is examined in [23]. A multimodal (text and picture) sentiment classification model based on a gated attention mechanism is provided in [24]. In the latter, the attention mechanism uses the image feature to highlight the text segment, allowing the machine to concentrate on the text that influences the sentiment polarity. Furthermore, the gating method allows the model to ignore the noise created during the fusion of picture and text, retaining valuable image information.

More examples can be found in [33, 34, 35] (see also Table 3.1) and in the very recent work presented in [36], which presents the Controllable Multimodal Feedback Synthesis (CMFeed) dataset, that enables, according to the authors, the generation of sentiment-controlled feedback from multimodal inputs. The dataset contains images, text, human comments, comments' metadata and sentiment labels. The authors propose a benchmark feedback synthesis system comprising encoder, decoder, and controllability modules. It employs transformer and Faster R-CNN networks to extract features and generate sentiment-specific feedback, achieving a sentiment classification accuracy of 77.23%.

Focusing on the large language model (LLM), other recent models exist. For example, in [37] the authors use transformers and LLM for sentiment analysis of foreign languages (Arabic, Chinese, etc.) by translating them into a base language - English. The authors start by using the translation models LibreTranslate and Google Translate, and the resulting sentences were then analyzed for sentiment using an ensemble of pre-trained sentiment analysis models, like Twitter-Roberta-Base-Sentiment-Latest, Bert-base-multilingual-uncased-sentiment, and Generative Pre-trained Transformer 3 (GPT-3). A 2024 survey about LLM and multimodal sentiment analysis can be found in [38].

Furthermore, recent multimodal methods, such as CLIP (Contrastive Language-Image Pre-training), BLIP (Bootstrapping Language-Image Pre-training), and VisualBert (see for

instance [39] or [40] for details), achieve excellent results in handling multimodal data. Nevertheless, some studies, like the one from Mao et al. [40], also suggest that using different pre-trained models for prompt-based sentiment analysis introduces biases, which can impact the performance of the model. Deng et al. [39] also addressed those models (CLIP, BLIP, and VisualBERT), validating that they are excellent models, but mentioning also as drawbacks they frequently have a lot of parameters and need image-text pairings as inputs, which limits their adaptability. The latter authors, Deng et al., implemented MuAL, which utilizes pre-trained models as encoders. A cross-modal attention is used to extract and fuse sentiment information embedded in both images and text with a difference loss incorporated into the model, to discern similar samples in the embedding space. Finally, a token (cls) was introduced preceding each modality's data to represent overall sentiment information. In a vision language pre-training model based on cross-attention (VLPCA) [41], a multi-head cross-attention to capture both textual and visual elements was used to improve the representation of visual-language interactions. In addition, to improve the performance of the model, the author created two subtasks and suggested a new unsupervised joint training strategy based on contrastive learning.

It is crucial to emphasize that there are a huge number of models available for text analysis [8], with the typical steps being as follows:

(i) *Processing text*: Using methods like tokenization, stop word removal, stemming, lemmatization, emoticon and emoji conversion, and deleting superfluous material, resulting in a "clean" text to improve the sentiment prediction accuracy.

(ii) *Extraction of Features*: Words, in this context, are relevant features that are extracted from the pre-processed text. The most popular method for doing this is to use strategies like n-grams and bag-of-words.

(iii) *Model Development*: Develop a DL model or a machine learning model (such as a Random Forest or Decision Tree) to learn from data and subsequently accurately classify newly unknown text sentiment.

(iv) Sentiment Classification Evaluation occurs when sentiment analysis is performed. To achieve the highest performance possible, the combination of several models and model adjustments are also common usage.

Table 3.1 summarizes different approaches/models to multimodal sentiment analysis. In parentheses is a small citation that points out the model. As can be seen, most of the models use different datasets, but all consider that text and images always have the "same" sentiment,

not considering that the text can be a specific sentiment, and the image can be a different one, or is only there to illustrate the post. Saying that, some of the models process the image and text separately, but in the end, join both (text and image) without any added consideration.

Also, it is possible to verify that, despite not being comparable, once there are different datasets and or different ways to use the (sub-)datasets, the accuracy (accr.) of the models is around 70% to 80%. As a note, in the last model presented in the table, the authors do not present the accuracy, only precision (P) and recall (R).

In the present work, ensemble/stacking modelling is suggested, which often enables data mining and predictive analytics applications to become more accurate. The process of running two or more related but independent analytical models and then integrating the results into a single score or spread is known as ensemble modelling, or fusion, in this context. In this instance, we can associate various outcomes from various/supplementary models to maximize accuracy or to provide complementary data. You may find examples of these methods, e.g., in in [18] and [24]. The pre-processing procedures used on the data before analysis, as well as the sub-datasets generated to evaluate the model, will be briefly discussed in the next section.

Table 3.1. Summary of multimodal approaches and respective accuracy.

Model (brief text citation)	**Ref.**	**Year**	**Dataset**	**Type**	**Accr.**
Deep Model Fusion ("…reduces the text analysis dependency on this kind of classification giving more importance to the image content…")	[19]	2019	B-T4SA	Text-img.	60.42%
Gated Attention Fusion Network (GAFN) ("…image feature is used to emphasize the text segment by the attention mechanism…")	[24]	2022	Yelp restaurant review dataset	Text-img.	60.10%
Textual Visual Multimodal Fusion (TVMF) ("…explore the internal correlation between textual and visual features…"	[33]	2023	Assamese news corpus	Text-img.	67.46%

Hybrid Fusion Based on Information Relevance (HFIR) ("…mid-level representation extracted by a visual sentiment concept classifier is used to determine information relevance, with the integration of other features, including attended textual and visual features …")	[34]	2023	Authors' dataset	Text-img.	74.65%
Deep Multi-Level Attentive Network (DMLANet) ("…correlation between the image regions and semantics of the word by extracting the textual features related to the bi-attentive visual features…")	[35]	2023	MVSA multiple Flickr	Text-img.	77.89% 89.30%
Transformer and Faster R-CNN Networks ("…controllable feedback synthesis to generate context-aware feedback aligned with the desired sentiment…")	[36]	2024	CMFeed	Text-img.	77.23%
Ensemble Model of Transformers and LLM ("…sentences were then analyzed for sentiment using an ensemble of pre-trained sentiment analysis models…")	[37]	2024	Compilation of several datasets	Text – "foreign languages"	86.71%
Multimodal sentiment analysis approach (MuAL) ("…cross-modal attention is used to integrate information from two modalities, and difference loss is utilized to minimize the gap between image and text information…")	[39]	2024	MVSA single, MVSA multiple, Hateful Memes, Twitter2015, Twitter2017	Text-img.	80.78% 77.77% 79.15% 79.34% 80.39%

Vision-Language Pre-training model based on cross-attention (VLPCA) ("…multi-head cross attention to capture textual and visual features for better representation of visual-language interactions…")	[41]	2024	Twitter2015, Twitter2017	Text-img.	(P & R) 71.20% 72.80%, 73.40% 74.00%

Lastly, to the best of our knowledge, no framework or model exists in the literature that considers the possibility that the post may elicit different reactions from readers to the image and text, regardless of the author's motivation for doing so. Also, assessing images of environments according to classes and integrating that data with the text seems to be missing from the literature. The next section describes the datasets used for the implementation of the proposed models and frameworks.

3.3 DATASETS

In the present study, three main groups of datasets were used: (sub-)datasets to develop and validate the Image Sentiment Classifier models (see Section 3.4.1), a dataset to build and validate the Text Sentiment Classifier models (see Section 3.4.2), and, finally, datasets used to develop and test the integrated model, the Multimodal Sentiment Classifier model (see Section 3.4.3).

3.3.1 IMAGE SENTIMENT DATASETS

In the image sentiment study, we used three image datasets, being one further divided into sub-datasets. In this context, from the (i) **Flickr dataset** (Flickr dataset available at: https://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html, accessed on 1 August 2024) [26], four sub-datasets are extracted. The reason is that the Flickr dataset includes outdoor, indoor, man-made scenes, scenes with persons etc., all mixed together, and it is intended to be used to train 4 distinct ISC models (see Section 3.4.1). To do this subdivision, a scene detection algorithm (code available at: https://github.com/AMANVerma28/Indoor-Outdoor-scene-classification, accessed on 1 August 2024) was used, which classifies whether an image is outdoor or indoor, and some

other attributes, such as man-made, natural light, no horizon, enclosed area etc. This algorithm, out of the focus of this dissertation, was used to speed up the segmentation of the initial dataset into the 4 sub-dataset. The resulting subsets of images were then visually validated by the authors. These sub-datasets were used to train and test the ISC models (see Section 3.4.1). Furthermore, although in Flickr's original dataset, the sentiment polarity of the images ranges from -2 to 2, the present models have only 3 sentiment classes: *positive*, *neutral*, and *negative*. This was achieved by defining a negative sentiment as between -2 and -1, a neutral sentiment as between -1 and 1, and a positive sentiment as between 1 and 2.

In more detail, the Flickr dataset was created from the computed Columbia Multilingual Visual Sentiment Ontology (MVSO) framework. This database covers a visual sentiment ontology consisting of 3,244 adjective-noun pairs and SentiBank, which is a set of 1,200 trained visual concept detectors, providing a mid-level representation of sentiment, associated training images acquired from Flickr, and a benchmark containing 603 photo tweets covering a diverse set of 21 topics. With ~461k images, the Flickr dataset is not balanced, so, ~179k balanced images were randomly selected from those 461k images (see Table 3.2, column "balanced"). From those, the images were subdivided into four sub-datasets (as mentioned above), one for each of the 4 environment categories: (i.1) **Flickr_ONMM**; (i.2) **Flickr_OMM**; (i.3) **Flickr_IND**; and (i.4) **Flickr_IOwPB**. As those sub-datasets were not balanced, for training and testing, when needed, they were balanced using data augmentation, namely: small rotations (5 degrees left/right), horizontal flips and zooms in (10% to 30%) on the images. Table 3.2 shows the number of images of the dataset as well as the sub-datasets and Figure 3.1 shows examples of images existing in the dataset.

Table 3.2. Summary of the Flickr sub-datasets for development of the Image Sentiment Classifiers.

Sentiment	**Flickr**		**Sub-datasets (Flickr)**			
	Original	**Balanced**	**ONMM**	**OMM**	**IND**	**IOwPB**
Positive (+)	280,157	**59,794**	32,878	28,653	12,664	57,639
Negative (−)	121,377	**59,794**	32,878	28,653	12,664	57,639
Neutral (=)	59,794	**59,794**	32,878	28,653	12,664	57,639
Total	461,328	**179,382**	98,634	85,959	37,992	172,917

Figure 3.1. Left to right, examples of images for the the four categories extracted from the Flickr dataset, i.e., ONMM, OMM, IND, and IOwPB. Top to bottom, examples of images with positive, neutral, and negative sentiment.

(ii) The **Simula Image Sentiment dataset (SIS)** (SIS dataset is available at: https://datasets.simula.no/image-sentiment/, accessed on 1 August 2024) [23] is a disaster-related dataset with ~3.7k images that was created by five different people. The ground truth classification was done by humans and varies between 1 (highly negative) and 9 (highly positive), with values between 1-3 being considered as negative sentiment, 4-6 as neutral sentiment, and 7-9 as positive sentiment. It is important to stress it is only a disaster-related dataset.

(iii) The **Image Sentiment Polarity dataset (ISP)** (ISP dataset is available at: https://data.world/crowdflower/image-sentiment-polarity, accessed on 1 August 2024) [42] contains over twelve thousand sentiment-scored images where the ground truth was at least approved by one human. There were five ground truth options: images classified as “Negative” or “Highly Negative” were associated with negative sentiment, “Neutral” with neutral sentiment, and “Positive” or “Highly Positive” with positive sentiment.

SIS and ISP datasets are used solely to infer the results of the ISC model, as they are the only datasets from the presented group that have been validated with human-verified ground truth. Being also important to emphasize that these two datasets are not balanced.

Table 3.3 summarizes the number of samples in the SIS and ISP datasets and sub-datasets, and Figure 3.2 shows examples of those samples.

Table 3.3. Sub-datasets are used for testing the inference of the ISC models.

Sentiment	**SIS & ISP**	**Sub-datasets (SIS & ISP)**			
	Original	**ONMM**	**OMM**	**IND**	**IOwPB**
Positive (+)	8,828	1,908	2,696	1,552	2,672
Negative (−)	3,996	242	1,327	581	1,846
Neutral (=)	2,963	273	951	709	1,030
Total	**15,787**	**2,423**	**2,842**	**4,974**	**5,548**

Figure 3.2. Left to right, examples of images for the four categories (ONMM, OMM, IND, and IOwPB), top to bottom, positive, neutral, and negative sentiment, for the SIS & ISP datasets.

3.3.2 TEXT SENTIMENT DATASET

For the text sentiment study, **the T4SA dataset** (T4SA is available at: http://www.t4sa.it/#dataset, accessed on 1 August 2024) [20] was used. The dataset contains ~3.4M tweets, corresponding to ~4M images, as each tweet may have more than one image. Each tweet, text and associated image(s) has been labelled according only to the sentiment polarity of the text, namely negative (-1), neutral (0), or positive (1). The dataset's authors removed corrupted and near-duplicate images and selected a balanced subset of images, named B-T4SA (more details in [29] and [20]). From that dataset, consisting of ~379k unbalanced

samples, a sub-dataset **B-T4SAtext**, which has 50k balanced samples was randomly selected for the TSC models (see Section 3.4.2). Table 3.4 shows the distribution of B-T4SAtext samples per sentiment category.

Table 3.4. Summary of the B-T4SAtext sub-datasets.

Sentiment	**B-T4SA**	**B-T4SAtext**
	Original	**Balanced**
Positive (+)	127,086	**16,667**
Negative (−)	21,643	**16,667**
Neutral (=)	203,471	**16,667**
Total	389,200	**50,001**

To this sub-dataset was applied the following pre-processing steps for text analysis: (a) the replacement of emojis/emoticons for words (e.g., 🥰 was replaced by "smiling face with hearts"); (b) convert all text to lowercase; (c) remove stop words (e.g., 'i', 'me', 'after', 'moreover'), which proves to be useful as stop words do not contribute to sentiment analysis and their exclusion avoids unnecessary computations; (d) removed HTML tags, images, mentions, links, punctuation etc., because they do not carry sentiments; (e) apply a lemmatizer, which removes inflectional endings from a token to turn it into the base word lemma (e.g., the word 'dancing' would be lemmatized to 'dance'); and (f) apply stemming, which is the process of removing suffices from words to obtain their root form (e.g., the word 'dancing' would be stemmed to 'danc'). Both stemming and lemmatizations serve the purpose of reducing word forms to their base or root forms to generalize the words, resulting in more accurate predictions in sentiment detection (see also [29]). Table 3.5 (the same as Table 2.2) shows examples of some texts from the **B-T4SAtext** dataset, alongside the same texts pre-processed according to the previously explained method.

3.3.3 MULTIMODAL SENTIMENT DATASET (IMAGE + TEXT)

The **B-T4SAmultimodal** sub-dataset was extracted from the T4SA dataset [20], being used for the MSC model (see Section 3.4.3). For this dataset, 1000 text posts have been randomly selected, carefully chosen by sentiment class, to ensure that the text samples are balanced. The text sentiment label was directly retrieved from the B-T4SA dataset, but the T4SA did not provide sentiment labelling (ground truth) for the images.

Table 3.5. Examples of preprocessed text used for the model's development (the same as Table 2.2). On the left is the sentiment, in the middle is the post text, and on the right is the pre-processed text used as input for the TSC model.

Sentiment	Text	PreProcessed Text
Negative (−)	RT @Reuters: Special Report: Iraq militia massacre worse than U.S. acknowledged. https://t.co/KhzqwloPam https://t.co/8HtpZYQZQh" (accessed on 1 August 2024)	special report iraq militia massacr wors u acknowledg
Positive (+)	Have a look at the most spectacular #NationalParks in #California. #traveler https://t.co/wiIfw7nnH7 https://t.co/sxnJnQH0cC (accessed on 1 August 2024)	look spectacular nationalpark california travel
Neutral (=)	Live Next Door To Parents https://t.co/FfJrf5n8D1 https://t.co/2bZRdnkPQY (accessed on 1 August 2024)	live next door parent

From each post, a single image *per* post was selected, presented and classified by a group of 10 persons, 6 male and 4 female, aged between 20 and 53 years, all with Portuguese nationality. The dataset then was filtered to include only those cases where a minimum of 6 out of the 10 individuals unanimously agreed on the sentiment classification (positive, negative, or neutral). This approach led to an unbalanced text and image (sub-)dataset, see Table 3.6, column "Used". This imbalance was due not only to the filtering process but also to other factors, such as the diversity of images/text sentiments pairing. For example, images classified as positive, neutral, or negative were sometimes paired with texts that had different sentiment classifications (negative, neutral, or positive). Consequently, after this image classification, from the initial 1000 posts samples, only 627 text – image samples remained. Table 3.6 shows the distribution of samples *per* sentiment used for training and testing the *Multimodal Sentiment Classifier* model.

In summary, the sub-datasets Flickr_ONMM, Flickr_OMM, Flickr_IND, and Flickr_IOwPB were used to train and test the ISC models; the SIS&ISP_ONMM, SIS&ISP _OMM, SIS&ISP _IND, and SIS&ISP _IOwPB were used only to test the ISC models; the (sub-)dataset B-T4SAtext was used to train and test the TSC models; and the B-T4Smultimodal_text, B-T4Smultimodal_ONMM, B-T4Smultimodal_OMM, B-T4Smultimodal_IND, and B-T4Smultimodal_IOwPB to train and test the MSC models. In the next section the Multimodal Sentiment Classification Framework and respective sub-modules are presented.

Table 3.6. Summary of the B-T4Smultimodal sub-dataset.

Sentiment	***B-T4SA ("TSC")***	***B-T4Smultimodal***					
		text		***images***			
	Original	Pre-classification	**Used**	**ONMM**	**OMM**	**IND**	**IOwPB**
Positive (+)	127,086	333	195	200	117	53	25
Negative (−)	21,643	334	177	18	45	19	20
Neutral (=)	203,471	333	255	3	26	84	17
Total	**389,200**	1000	**627**	**221**	**188**	**156**	**62**

3.4 MULTIMODAL SENTIMENT CLASSIFICATION FRAMEWORK

As already mentioned, the *Multimodal Sentiment Classification model* (MSC) is used to extract sentiments from images and texts, following the principles presented in [29] and [19]. However, the present model exhibits distinctions, while [29], the previous work of the present authors, focused only on posts (text-image) associated with landscapes. In [19], the authors achieved 60.42% accuracy on a test set of 51k samples from a B-T4SA image-balanced sub-dataset, with the currently considered three classes (negative, neutral, and positive). Nevertheless, in [19], the authors did not consider that a post can have more than one associated image and the fact that image and text can reflect the same or different sentiments. Additionally, a post with multiple images may present different sentiments, which may or may not coincide with the text sentiment. The present work focuses on improving accuracy, but also, most importantly on differentiating/marking posts images and texts that reflect different sentiments.

In these works, ensembles are employed to improve the framework's accuracy. Namely, image and text sentiment classification results from the combining of various methods, i.e., components are combined into an ensemble to yield the ultimate result.

In this context, the possible outputs are (see Figure 3.3): (i) the sentiment (*isc*) resulting from the image (ISC), generated by the *Image Sentiment Classification* block; (ii) the sentiment (*tsc*) resulting from the text (TSC), generated by the *Text Sentiment Classification* block; the (iii) sentiment (*msc*) resulting from the combination of image and text (MSC), generated by the *Multimodal Sentiment Classifier* block; and (iv) the discrepancy (*dis*.) between image and text. So, for each pair (text -image), the model returns the following sentiment classifier vector SCv = [*isc*, *tsc*, *msc*, *dis*.].

Figure 3.3 shows a simplified diagram block of the model, where "+" represents a positive sentiment, "-" a negative sentiment, "=" the neutral sentiment, and dis. is the discrepancy between the sentiment returned from the image and the text, for each pair of text-image presented in the input.

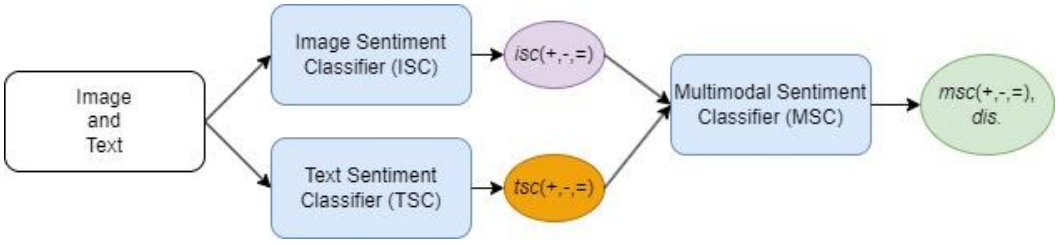

Figure 3.3. Multimodal Sentiment Classification Framework.

3.4.1 IMAGE SENTIMENT CLASSIFICATION

The *Image Sentiment Classification* model combines several individual image sentiment classifier models into an ensemble that predicts the final sentiment. In this context, see Figure 3.4 top, the ISC is made up of four blocks that correspond to each class of sentiment classifier, i.e., the ISC is the result of the ensemble of the results of the ISC for *non-man-made outdoors* (ISC_ONMM), for *man-made outdoors* (ISC_OMM), for *indoor* (ISC_IND), and for *indoor/outdoor with persons in the background* (ISC_IOwPB).

Each sentiment classifier, ISC_*class*, with *class* ∈ {ONMM; OMM; IND; IOwPB}, returns the probabilities of an image carrying a negative, neutral, or positive sentiment. The outputs of the four class sentiment classifier blocks are then ensembled using a Random Forest (ISC_RF) and a Neural Network (ISC_NN). Each of these individual blocks, ISC_*class*, return results for each category (*class*), again doing an ensemble of the three DL models that will be presented below, this is illustrated in Figure 3.4 bottom. This last step follows the author's previous work done for ONMM [29].

In summary, ISC_RF or ISC_NN fuses the 36 probabilities given by the 4 blocks (9 answers for each block resulting from the 3 probabilities outputs by each of the 3 individual models) to finally decide the final sentiment of an image (ISC). The number of individual models is a hyperparameter that can be tuned.

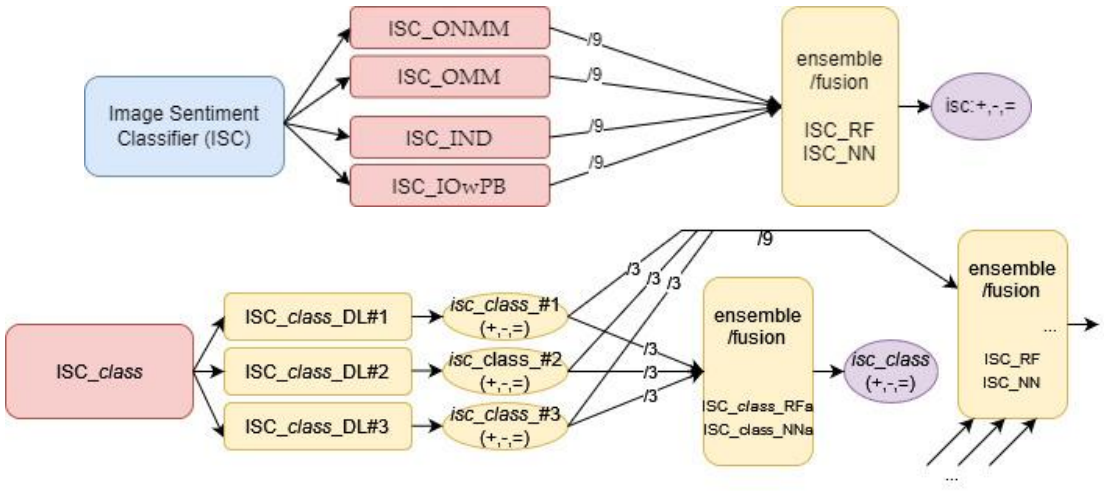

Figure 3.4. Top, the ISC model, bottom the ISC sub-block specification, with class in {ONMM; OMM; IND; IOPB}.

The above-mentioned 3 models, for each ISC_*class*, have DL architectures. The architectures are based on backbones from known architectures followed by a handcrafted network head, based on fully connected layers. In this scenario, three distinct backbones were used to extract different types of features, namely: EfficientNetB0 [26], Xception [27], and ResNet-50 [28]. It is worth noting that all backbones were trained using the well-known ImageNet dataset.

Different strategies of transfer learning were used for the models, including different numbers of fully connected layers at the networks' heads. Nevertheless, all ISC_*class* blocks have the same three individual models, with the same architecture, and hyperparameter, that were only optimized for the ONMM *class*. What differentiates an individual classifier from the others is the class category of the sub-dataset images (Flickr_ONMM, Flickr_OMM, Flickr_IND and Flickr_IOwPB) that were used to train the individual and ensemble models.

The deep learning models are:

(a) **Model DL#1:** the backbone is an EfficientNetB0 (237 layers). Furthermore, let $L_{i,j}$ represent a dense layer, where *i* is the layer number and *j* is the number of units. Then, in ISC_*class*_ DL#1, the head has 5 layers. In the initial layer, L_{in}, the number of units equals the number of outputs of the backbone. The second dense layer has *n* units with X_2 activation function ($L_{2,n}$), a dense layer $L_{3,m}$, and a dense layer with 24 units ($L_{4,24}$), all with X_i activation functions (see Table 3.7). The last layer has 3 units with a softmax activation function (Ls).

(b) **Model DL#2:** the backbone is Xception (71 layers), and the head has also 5 layers. In the ISC_*class*_DL#2, the first dense layer has L_{in} units, equal to the number of outputs of the backbone. Then, it has a dense layer with *n* units ($L_{2,n}$). The third layer has also *m* units and $L_{3,m}$. Then a dense layer with 24 units ($L_{4,24}$). All layers have X_i activation functions. The last layer is the Ls.

(c) **Model DL#3:** the backbone is ResNet-50 (50 Layers), and the head has 4 layers. In the ISC_*class*_DL#3, the first dense layer is L_{in}, the second layer $L_{2,n}$, the third layer $L_{3,24}$, all with X_i activation functions, and the last layer is the Ls.

These architectures were tuned using the sub-dataset ONMM. In this context, several hyperparameters were tested, such as the number of units, batch size, number of epochs etc. (see Section 3.5 for the hyperparameters). The only fixed values were the penultimate (with 24 units) and the last (with 3 units) layers, and the softmax activation function. The reason for the layer of 24 units is based on the authors' hypothesis derived from Plutchik's wheel of emotions. The authors hypothesize that due to the importance of color in sentiment analysis and the relation between emotion and sentiment, and once there are 24 emotions in Plutchik's wheel, it is expected that a layer of 24 units can help the network to learn those emotions and relate them with the 3 sentiments, which appears in the last layer. I.e., it will allow the image classification of the sentiment into positive, negative, or neutral (reason for 3 units in the last layer) according to the responses of those "emotions".

It should be noted that although there are only 3 models here, more than 100 models have been tested for *class* ONMM, with different hyperparameters, optimizations, and backbones. No drop-out layers were used in these 3 models either, but they were also tested.

In the next step, the results from the three models are ensembled to obtain a final classification. The inputs of the ensemble sub-block will be the predictions made by the individual models, resulting from the softmax function, and the output will be the final image sentiment prediction. Ensembling leverages the idea that combining the strengths of multiple models can result in improved overall performance and accuracy, compared to using a single model, as well as a better generalization (reacts better to unseen data than single models), reduces "overfitting", and increases the robustness of the results obtained. For the aggregation of sentiment classification, the ensemble sub-block, was used:

(i) ***Random Forest*** **(ISC_RFa)**: *k* estimators (see Section 3.5), Gini impurity as criteria function to measure the quality of split, and the minimum number of samples required to split an internal node was 2. The rest of the hyperparameters were set to the default

values of the scikit-learn library (https://scikit-learn.org/, accessed on 1 August 2024) (v. 1.5).

(ii) ***Neural Network*** **(ISC_NNa)**: three dense layers, where the first layer has 9 units (L_{in}), then a n layer with m units ($L_{n,m}$) with Xi activation functions, and the third layer the Ls. A search tuner function was used to find the best (hyper-)parameters for the proposed model. The only layer not found by the search tuner is the last, Ls layer, of 3 neurons, which uses the softmax activation function to obtain the probability of sentiment.

3.4.2 TEXT SENTIMENT CLASSIFICATION

In the *Text Sentiment Classification* block, the first step consists of converting the text (see Section 3.3.2) into structured data, in a way that can be used by ML methods. To achieve this, a Bag of Words was applied to extract 5.000 features, with one to two n-grams, from the processed texts, see details in [29], and the ML models used are:

(a) Random Forest (TSC_RFc): created with *k* estimators and Gini impurity as criteria function to measure the quality of the splits**.** The rest of the hyperparameters were set to the default values of the scikit-learn library (v. 1.5).

(b) Neural Network (TSC_NNc): 4 dense layers where the first has 5.000 units (L_{in}), then two layers with *n* units ($L_{2,n}$) units and *m* units ($L_{3,m}$) are added, with X_i activation functions. The last layer uses 3 units with a softmax activation function (Ls), used to predict the probability of a text carrying a positive, negative, or neutral sentiment.

(c) Natural Language Toolkit (TSC_NLTK) method was used as a third model. (NLTK available at: https://www.nltk.org/, accessed on 1 August 2024)

The block diagram of the proposed TSC is presented in Figure 3.5. For the aggregation of text sentiment classification (the ensemble sub-block) was used:

(i) ***Random Forest (TSC_RFt)*****:** *k* estimators and "Gini criteria" function (as before). The rest of the hyperparameters were set to the default values of the scikit-learn library (v. 1.5);

(ii) ***Neural Network (TSC_NNt):*** 4 dense layers, the first layer has 9 units (L_{in}), then a dense layer with 3 units ($L_{2,n}$), a third layer with m units ($L_{3,m}$) all with Xi activation function, and the fourth layer the Ls.

As mentioned, a search tuner function was used to find the best (hyper-)parameters for the proposed models (see Section 3.5).

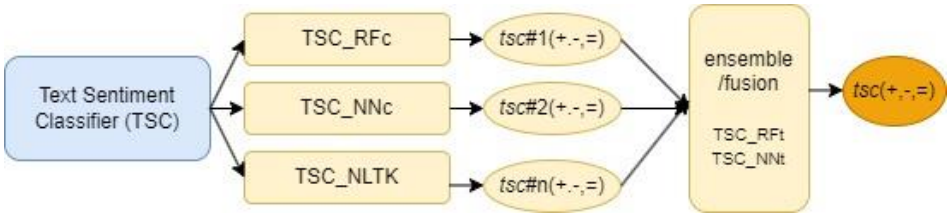

Figure 3.5. Block diagram of the Text Sentiment Classifier.

3.4.3 MULTIMODAL SENTIMENT CLASSIFICATION

The *Multimodal Sentiment Classifier* block, as largely mentioned in the text, is based on classifications resulting from text and image. Going to the block diagram in Figure 3.6, from the ISC model there are 36 output probabilities (4 *classes* × 3 individual models × 3 sentiments), from which are used the 9 output probabilities (3 individual models × 3 sentiments) from the individual models corresponding to the *class* of the image that is being analysed. From the TSC are also used the 9 output probabilities (3 individual models × 3 sentiments). Also, *isc* and *tsc* sentiment classification are used to compute the *discrepancy*, which acts as a selector to compute the *msc* based on the ISC and TSC, as we will see next.

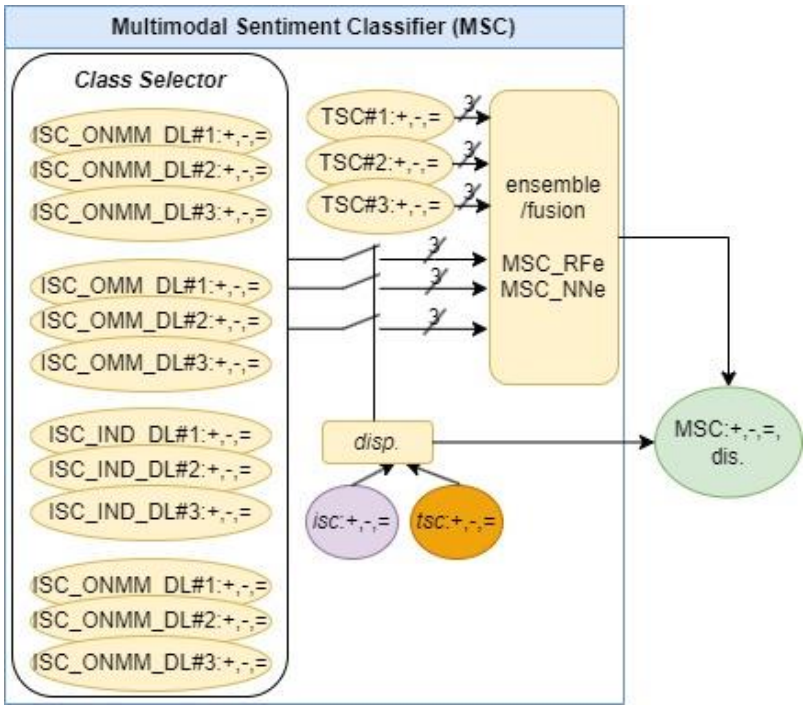

Figure 3.6. Block diagram of the Multimodal Sentiment Classifier.

The *discrepancy (dis.*) varies between 0 and 100, with 0 corresponding to the sentiment between the image and text being the same, and 100 for the sentiment between the image and text being different, and it is computed as follows:

(a) Similar sentiments, i.e., *isc = tsc*, then *msc = isc = tsc*, and SCv = [*tsc*, *tsc*, *tsc*, *dis. = 0*] is the resulting output.

(b) Hypothetic different sentiments, i.e., $isc \neq tsc$, then considering μ_I the average results (between 0-100) from all ISC models that returned the predicted sentiment, μ_T the average of all TSC models that returned the predicted sentiment, and a threshold μ_t = 85 (empirically calculated), two cases can occur:

(i) ***The image and text have clearly different sentiments:*** If $\mu_I \geq \mu_t$ and $\mu_T \geq \mu_t$ then both ISC and TSC are considered to be certain of the predicted sentiment and, in this case, most probably the person who posted the text-image "intended" different sentiments. The ensemble block is not computed, resulting SCv = [*isc*, *tsc*, ×, *dis.* = 100].

(ii) ***The image and text have indeterminate sentiments***: The remaining cases. It means that the ensemble between text sentiments and image sentiments must be computed, once there is no certainty about what the person intended to post. Resulting in SCv = [*isc*, *tsc*, *msc*, *dis.*], where the $dis. = 100 - (\mu_I + \mu_T)/2$ if *tsc < 50* or *isc < 50*, or $dis. = (\mu_I + \mu_T)/2$ for the remaining situations.

As there is no balance in the sentiment classes dataset and the number of samples is relatively small, once the B-T4Smultimodal (sub-)dataset is used, the Stratified K-Fold cross-validation technique is used to train the MSC model (K=5). The ensemble block is computed following the same principles presented before, namely:

(i) ***Random Forest (MSC_RFe)*:** *k* estimators, *l* minimum samples in a leaf, *m* minimum samples required to split the internal node and Gini as criteria function. The rest of the hyperparameters were set to the default values of the scikit-learn library (v. 1.5);

(ii) ***Neural Network (MSC_NNe)*:** 4 dense layers, where the first layer has 18 units (L_{in}), a dense layer with *n* units ($L_{2,n}$), a third layer with *m* units ($L_{3,m}$), with X_i activation function, and the fourth layer the Ls with softmax as activation function.

The following section details the tests conducted and their results, followed by a discussion of the work undertaken.

3.5 TESTS, RESULTS AND DISCUSSION

The tests and discussion were divided into 3 sections, one dedicated to the ISC, the other to the TSC (showing the results achieved in [29], to better understand the present chapter), and the final combination of text-image (post), the MSC.

3.5.1 IMAGE SENTIMENT CLASSIFICATION

The procedure implemented to evaluate ISC_*class* (individual) models is divided into two steps: (a) evaluate image sentiment classification on test data of Flickr sub-datasets and (b) evaluate the models in the SIS and ISP datasets. The models were trained and tested using the Kaggle's platform (with 14.8 GB RAM e 2x GPU NVIDIA T4).

All ISC_*class* (individual) models were trained with 70% of the samples, being from the remaining 10% to validate the model, and 10% for testing. For the ensemble model evaluation, 30% of the previous samples were considered, the ones not used for training the individual (*class*) models (namely, 10% from the validation, 10% from the test, and 10% remaining unused data) were used. From these 30%, 80% of the samples were used for ensemble training and 20% for testing.

To validate the hypothesis that 4 ISC models, one *per class*, work better than a single model trained with all images, a Holistic Image Sentiment Classifier (HISC) was trained using the Flickr sub-dataset listed as "balanced" in Table 3.2 (70% samples for training, 10% for validation, and 20% for testing),. The HISC uses the 3 DL blocks/models and the same previously stated ensemble strategies to predict the sentiment of an image. The difference lies in the training samples: ISC models divide the samples by classes, whereas HISC uses all samples together.

Table 3.7 shows the model's hyperparameters. It is important to emphasize that the classifier models for each *class* and the HISC use the same hyperparameters. In more detail, the first column of the table shows the used models, and the second column summarizes the number of units used in the layers (resulting from the optimizations), as well as the number of estimators. The remaining columns present how many layers the backbone was trained with, as well as the hyperparameters used to train with the new data. There is only one exception: for the ensemble models that were built using a neural network, a search tuner was used to choose the network settings, and hyperparameters, to obtain the best possible result.

For the Random Forest ensemble models, the Gini impurity was the criteria function to measure the quality of split and the minimum number of samples required to split an internal

node was 2 (the rest of the hyperparameters were set to the default values of the scikit-learn library). The model was initially trained in two ways: (i) by using the sentiment values of -1 (negative), 0 (neutral), and 1 (positive), predicted by the individual models; and (ii) by using the sentiment-predicted probabilities from the individual models. After analysis, option (ii) was chosen, because the results indicated a significant improvement when comparing to option (i), as we can see in Figure 3.7. The figure shows ISC_OMM models' confusion matrices. On the left, the model uses 3 inputs (option i) from the direct sentiments of the individual models, achieving an accuracy of 80.81%, and on the right the model uses 9 inputs corresponding to the probabilities of the sentiment (ii), achieving an accuracy of 82.21%.

Table 3.7. ISC_class, HISC, and ISC individual and ensemble models backbones and hyperparameters.

Model	**Number units / estimators**	**Backbone (layers trained)**	**Hyperparameters**	**Activation Function**
Model DL#1	n = 1024 m = 512	EfficientNetB0 (none)	Opt: Adam (1e-4) Epochs: 20 Batch size: 32	Xi = ReLU i={1,…,5}
Model DL#2	n = 1024 m = 512	Xception (8)	Opt: Adam (1e-4) Epochs: 20 Batch size: 32	Xi = ReLU i={1,…,5}
Model DL#3	n = 512	ResNet50 (12)	Opt: Adam (1e-4) Epochs: 20 Batch size: 32	Xi = ReLU i={1,…,4}
RFa	k = 100 (est.)	-	-	-
NNa	decided by the optimizer	-	Batch size: 32	-

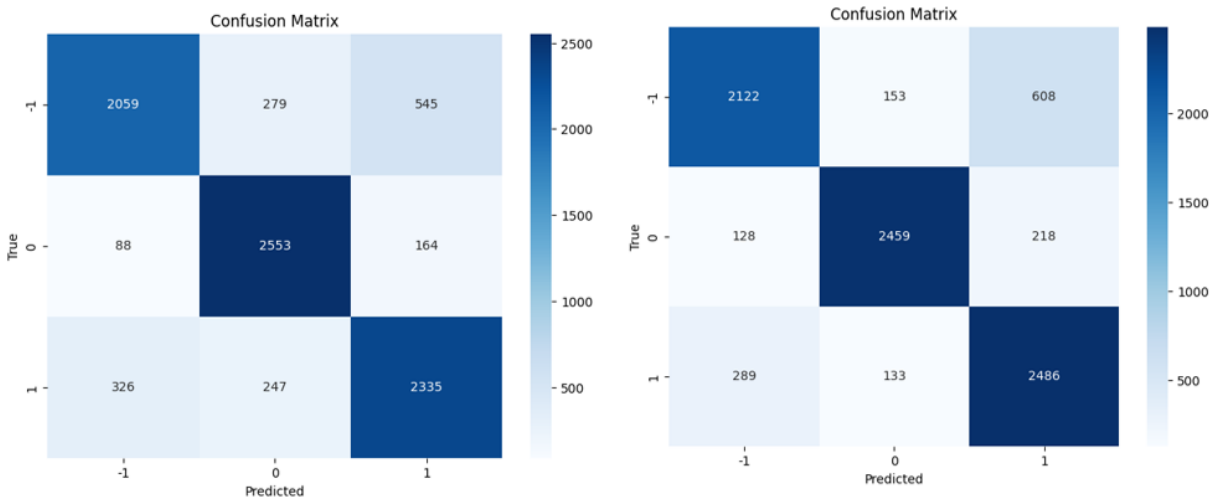

Figure 3.7. ISC_OMM models' confusion matrices: on the left, the model uses 3 inputs (which are the direct sentiments of the individual models), while the model on the right uses 9 inputs (corresponding to the probabilities of the predicted sentiment).

The accuracies for the different sub-datasets and ensemble are presented in Table 3.8, where the first column indicates the class, the second the model used, and the remaining columns the accuracy for the Flickr sub-datasets and SIS & ISP sub-datasets. When evaluating the models, the ensemble of individual models (ISC) presented better results than the holistic model (HISC) for the Flickr and SIS & ISP sub-datasets, as marked in grey in the table. For Flickr sub-datasets, ISC achieved 76.45% compared to HISC's 65.97%. Similarly, for the SIS & ISP sub-dataset, ISC achieved 53.31% compared to HISC's 47.80%.

Going down to the individual models, the results for the Flickr sub-datasets are all above 60%, except for two cases of model DL#3. For the SIS & ISP datasets, the results are less favorable, as there is some accuracy below 50%, but always above 44%. It is important to remember that humans did the labelling of SIS & ISP, unlike Flickr. For the Flickr sub-datasets, the individual ISC models achieved accuracies of 84,54%, 83.21%, 68.53%, and 84.79% for the classes ONMM, OMM, IND, and IOwPB, respectively. For SIS & ISP, the accuracies were 61.53%, 56.92%, 51,62%, and 49,06% for the same classes, respectively.

Table 3.8. ISC_class, ISC, and HISC individual and ensemble models accuracy.

Class	***Model***	***Flickr sub-datasets accuracy***	***SIS & ISP accuracy***
ISC_ONMM	Model DL#1	80.30%	53.82%
	Model DL#2	77.87%	53.61%
	Model DL#3	71.09%	50.23%
	RFa	**84.54%**	61.16%
	NNa	83.34%	**61.53%**
ISC_OMM	Model DL#1	79.31%	50.36%
	Model DL#2	76.29%	53.40%
	Model DL#3	55.26%	47.29%
	RFa	**83.21%**	55.93%
	NNa	83.09%	**56.92%**
ISC_IND	Model DL#1	64.53%	46.31%
	Model DL#2	61.37%	49.65%
	Model DL#3	55.47%	42.93%
	RFa	**68.53%**	50.21%
	NNa	67.47%	**51.62%**
ISC_IOwPB	Model DL#1	81.46%	44.85%
	Model DL#2	77.65%	46.99%
	Model DL#3	64.75%	46.41%
	RFa	84.79%	**49.06%**
	NNa	84.52%	48.77%
HISC	Model DL#1	63.46%	**45.25%**
	Model DL#2	61.85%	**44.31%**
	Model DL#3	54.74%	**41.97%**
	RFa	**65.97%**	47.21%
	NNa	64.45%	47.80%

ISC	RFa	**76.45%**	**53.31%**
	NNa	69.92%	48.51%

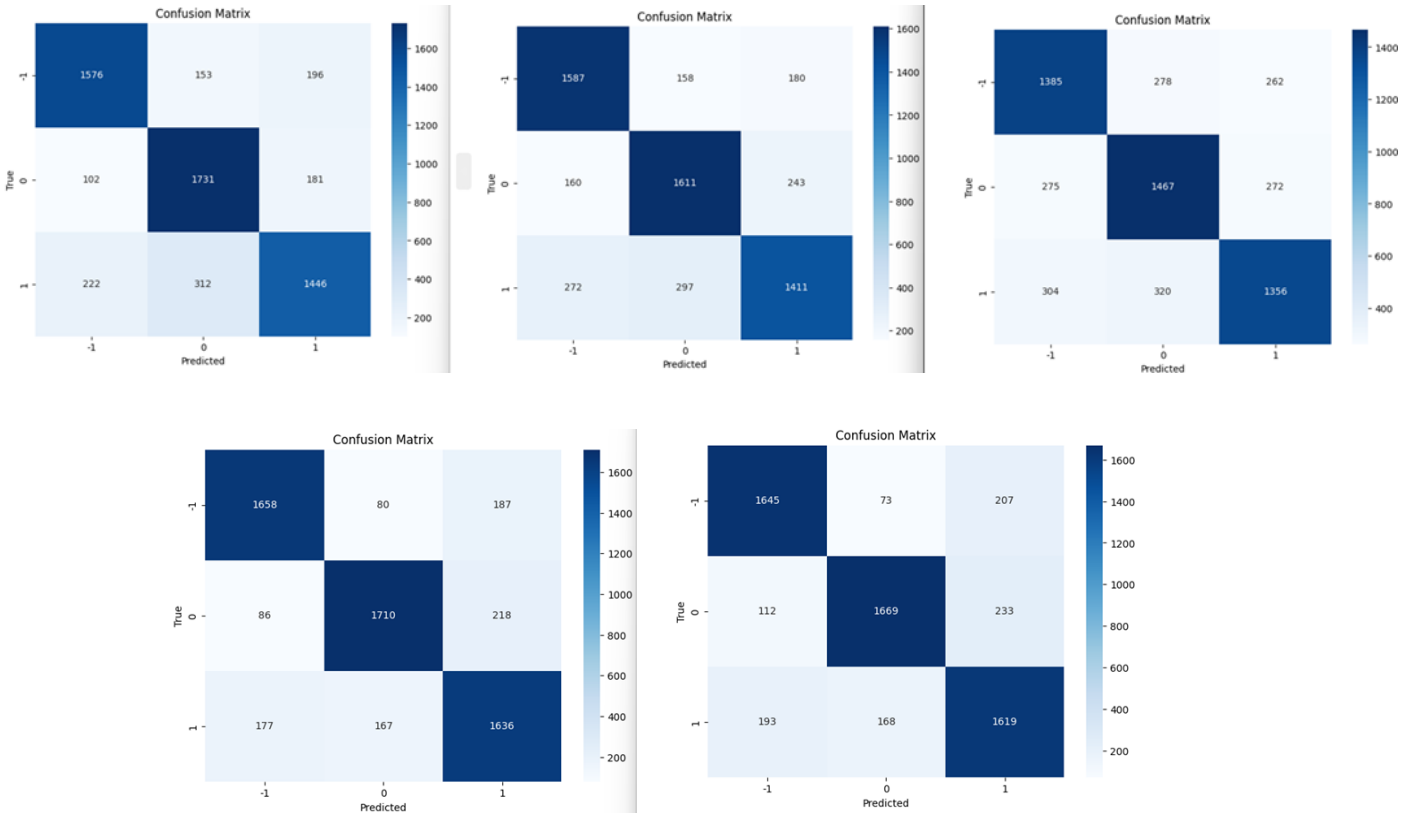

Figure 3.8. ISC_ONMM confusion matrices for models DL#1, DL#2, and DL#3 (top line, from left to right), and ensembles RFa and NNa (bottom line, left to right).

Another strong indicator is that the test accuracies are very close for the three individual models *per class*, as this consistency suggests that they are all effectively capturing the underlying patterns in the data, leading to somehow similar performance. Additionally, the ensemble models play an important role as they enhance the accuracy of the best individual models by ~4% to ~5%, consistently in all classes. The same occurs for the global models. In addition, the ensembles offer stability in their responses as they are based on the multiple analyses of the other (individual) models. Figure 3.8 highlights the close test accuracies of the individual models, using the example of ISC_ONMM confusion matrices for models DL#1, DL#2, and DL#3 (top line, from left to right), and ensembles RFa and NNa (bottom line, left to right).

Once again it is important to stress that the same backbones, hyperparameters, and parameters (configurations) were used for all individual models and ensembles, ISC_*class*, ISC, and HISC model. Fine-tuning the models will achieve (for sure) better performance

(accuracy); however, that was not the goal of this study, as the aim was to compare the models using consistent configurations.

3.5.2 TEXT SENTIMENT CLASSIFICATION

For the development of the individual and ensemble TSC models, the Colab platform was used with 12.7GB of RAM and 107.7GB of disk space. Table 3.9 shows the configurations and accuracy of the individual and ensemble models for TSC that were trained using B-T4SAtext dataset. It also displays the accuracy of the individual models while highlighting the effectiveness of ensemble models in stabilizing and enhancing their accuracy. For more details about this section please see [29]. A final observation, it is important to highlight that the TSC model, utilizing neural networks, achieves the best result with an accuracy of 92.10%.

Table 3.9. TSC parameters, hyperparameters, and accuracy of the models.

Model	**Number units /estimators**	**Hyperparameters + activation function**		**Accuracy**
TSC#1 (RF)	k = 100 (est.)	—	—	90.10%
TSC#2 (NN)	n = 300 m = 100	Opt: SGD (1e-2) Epochs: 10 Batch size: 8	X_i = ReLU i={1,…, 3}	88.55%
TSC#3 (NLTK)	—	—	—	84.34%
TSC – RFt	k = 100 (est.)	—	—	91.95%
TSC – NNt	n = 100 m = 20	Opt: SGD (1e-2) Epochs: 10 Batch size: 2	X_i = ReLU i={1,…, 3}	**92.10%**

3.5.3 MULTIMODAL SENTIMENT CLASSIFICATION

The MSC was trained in the Kaggle's platform (with 14.8 GB RAM e 2x GPU NVIDIA T4) using B-T4SAmultimodal and a stratified K-fold cross-validation technique, focusing only on samples where the image and the text share the same sentiment (273 samples). Before entering more details about MSC results, characterizing the samples per *discrepancy* is important.

From the B-T4SAmultimodal dataset (classified by humans), the MSC framework using the specification stated in Section 3.4.3, separates samples into three categories: those where the image and text represent the same sentiment (38.60%), those where there is uncertainty about their alignment (55.66%), and those where the sentiments are clearly different (see Table 3.10).

This means that 38,60% (242 from 627 samples) plus 5.74% (36 from 627 samples) are already classified (no additional processing is required). The former because *isc* equals *tsc*,

and consequently *msc* equals *tsc* returning SCv = [*tsc*, *tsc*, *tsc*, *0*]; the latter because there is no possible classification for *msc*, once the model is completely sure of the sentiment for ISC and for TSC, and they are different, meaning SCv = [*isc*, *tsc*, ×, 100]. The framework only needs to do the inference for the samples that have $0 < dis. < 100$, in the present dataset 55.66% (i.e., 349 from the 627 samples).

Table 3.10. Number of samples per sentiment discrepancy.

Discrepancy	**Number of samples**	**Percentage of samples**
dis. = 0	242	38.60%
0 < *dis.* < 100	**349**	**55.66%**
dis. =100	36	5.74%

Table 3.11 shows the parameters, hyperparameters, and accuracies *per* discrepancy of the MSC model. In more details, if *dis.* = 0 (ISC and TSC report the same sentiment) then the framework reports a 100% accuracy, i.e., these samples need no further computation, once *isc* equals *tsc*, so it just checks the sentiments against the ground truth (no inference was done). For $0 < dis. < 100$, the best result was achieved for the ensemble using random forest, with an accuracy of 64.18% (using the neural network, the result is similar with a difference ~4%).

At this point, it is again important to stress how the accuracy of MSC is computed, as it aggregates results from the ISC model and TSC model, and checks these predictions against the ground truth, once the image and text reflect or could reflect different sentiments. Following this, the good results in the accuracy of the MSC model (above 78%) are justified by the fact that: (i) the final accuracy is computed between the samples for *dis*. = 0 (100% accuracy; no ensemble model was applied) and the samples from $0 < dis. < 100$ (64.18% accuracy), which returns the final accuracy of the model of 78.84%. (ii) They are exclusively employed to determine and enhance sentiment accuracy based on the results of the 12 individual models (3 TSC and 3 ISC_*class*). (iii) For the remaining samples, when *dis.* = 100, the ensemble is not applied, once it is not computed the *msc*, i.e., SCv = [*isc*, *tsc*, ×, 100].

Table 3.11. MSC parameters, hyperparameters, and accuracy.

Model MSC		**Number units /estimators**	**Hyperparameters + activation function**		**Accuracy (samples)**
MSC_RFe	*dis*. = 0	—			100.00% (242/242)
	0 < *dis*. < 100	*k* = 150 (est.) *l* = 4 (samples) *m* = 5 (samples)	—	—	64.18% (224/349)
MSC Model (using Random Forest - MSC_RFe)					**78.84% (446/591)**
MSC_NNe	*dis*. = 0	—			100.00% (242/242)
	0 < *dis*. < 100	*n* = 466 *m* = 24	Opt: Adam(7.88e-4) Epochs: 20 Batch size: 4	X_i = ReLU *i*={1,…, 3}	60.17% (210/349)
MSC Model (using Neural Network - MSC_NNe)					76.48% (452/591)
N.A.	*dis*. = 100	—			Not applicable

It is important to note that the number of samples existing in B-T4SAmultimodal to train and test the MSC model is very low, and this can bias the results. Nevertheless, the results prove that is possible to detect if an image and text share the same sentiment, as well as to identify instances where they have completely different sentiments. Furthermore, the framework effectively combines images and text that may or may not have different sentiments, achieving good result when comparing with the classification done by humans.

Despite this, it is mandatory for future works to increase exponentially the number of samples classified by humans (a second version of the B-T4SAmultimodal) and make it public for other authors to test the results against these initial baseline results. The present dataset will be available at https://osf.io/institutions/ualg (accessed on 1 August 2024).

One of the difficulties of this research is the nonexistence of a sentiment dataset classified by humans that can validate the main research goals. Specifically, these objectives are: (i) developing a single hyperparameter image classification sentiment model that performs well across different environments, including validating that the classification accuracy supported on 4 categories is better than using a single category (including all images); and (ii) the development of a framework that combines image and text sentiment classification. The framework must return a multimodal sentiment classification along with the discrepancy metric, which includes the idea that text and image can only be combined if both return the same sentiment or if both return uncertain sentiments.

In the latter case, one text/image can complement the other to achieve the final sentiment classification. When both return different sentiments, they should not be joined. I.e., when conflicting sentiments occur, empirically, is possible to say that the person posted the text with a sentiment and used the image only for illustration purposes, or the opposite, posted the image with sentiment and the text is only to "frame" the image. This needs further research, including how these posts can/should be used by managers who receive this information to manage their companies or platforms.

Returning to the non-existence of a dataset with ground-truth sentiment for text and images from posts, the solution adopted in this work to mitigate this problem was to use multiple datasets and sub-datasets. While this is not an ideal solution, it effectively supports the scientific goals of the paper.

For the TSC it was used a single dataset that filled all the requirements. For the ISC, the initial dataset, classified automatically, was divided into 5 classes. From these 5 classes, the one depicting in the foreground (semi-)frontal humans was discarded due to 3 main reasons: (i) this class is certainly the most analyzed in the literature, presenting excellent models that validate the sentiment, most of the cases using the human face; (ii) text and image sentiment usually, in this case, are very similar; (iii) in a post indoor and outdoor scenes many times are used with different purposes, such as transmitting a sentiment, be ironic, illustration etc. Usually, images with faces transmit a specific emotion in posts. For the MSC it was not possible to find a dataset that had human-annotated ground-truth classification for the image, the text, and the combined image-text.

When developing the author's dataset, it was validated that these three possibilities are very crucial. For example, a person might classify a text with a positive sentiment and an image with a negative sentiment. However, when viewing both the image and the text together, the overall sentiment might be different, such as neutral. In reality, the last one is the one meaningful for (training) the MSC.

In consequence, despite the author's dataset having a limited number of samples, there were enough to validate the planted concept. In the author's dataset, 55.66% of the samples fall in the mentioned situations, which is not an insignificant number. In summary, there are (sub-)datasets that fill the requirements to validate the paper's goals. Nevertheless, for future work, all the datasets need more samples classified by humans. In addition, the feedback of the post sentiment to the managers who receive the information to manage their companies or platforms should be the image sentiment, the text sentiment, the combined (text-image)

sentiment if exists and the discrepancy between image and text sentiment (SCv = [*isc*, *tsc*, *msc*, *dis.*].

Also related to the dataset are the failures of the models, which are detected mostly in the ISC. The classification of the images where the ISC models detect positive sentiment and the image is negative or select negative and the image is positive (see Figure 3.7 and Figure 3.8) is mostly due to the datasets not being balanced and the difficulty of having a clear classification of "neutral" image. Figure 3.9 shows examples of images where different persons gave different classifications, going from the positive, to the neutral, to the negative. Take the example on the right; for someone who appreciates history, the knight's armor evokes a positive sentiment. For other, the complete scene might be perceived as neutral. However, for different individuals, the image might conjure thoughts of war and medieval times, leading to a negative sentiment. So, to develop a more robust model, the dataset must have not only the human classification, but also the characteristics of the human classifier, and the model should also account for that information. For more information about this subject please see [43].

Similar issues arise in the MSC due to the lack of a comprehensive dataset. When evaluating both text and images together, human classifiers sometimes base their decisions on group considerations or personal biases—such as a preference for images over text, or vice versa—leading to inconsistent classifications for the same text-image pair. The solution to this problem would be to have a large number of samples for training the classifier. However, such a dataset is not currently available and should be addressed in future work.

Figure 3.9. Examples of images where the human's classification presented more doubts.

All the above leads us to compare the current framework with state-of-art models. In terms of results, the present framework achieves 78.84% accuracy, which places it among the top results presented in Table 3.1. It is possible to say that the present framework is at the top results, or at least is in line with the state-of-the-art results. However, a direct comparison is not a completely fair for the present framework or for the models presented in the table, once: (i) all use different datasets or sub-datasets; (ii) some are trying to achieve the best result

possible, while the present framework is proving/showing/presenting a concept/idea/goal; (iii) only some models are trained with human-classified data; and (iv) some models train image and text classifiers separately and then combine them, without using data where both image and text are classified together by humans. Bottom line, at the moment the results in this specific area cannot be comparable until the authors from different publications use the same procedure. Nevertheless, as mentioned, it is possible to validate that the present results are completely in line with the best state-of-the-art results.

Examining the results of the individual models in detail, Table 3.8 and Table 3.9 present the results of the ISC and TSC individual models, and respective ensembles. It can be seen (as already discussed) that the ensembles return better results than the individual models (which was one of the research questions). Not mentioned, but also important to stress, each (individual) ISC model uses different backbones, and, consequently, extracts different features from the image. The same occurs for the text, i.e., TSC_RFc and TSC_NNc use the same features which are different from TSC_NLTK.

Another important point to reinforce is the training of the models. As mentioned, the ISC_ONMM was tuned using more than 100 variations of hyperparameters of the network's head. Each training procedure took around 4 hours, for each variation of parameter, in the Kaggle's platform. Having this classifier fine-tuned and considering this category the more generic one, it was considered (hypothesized) that the same parameters could be used for all the categories, reducing the time used for fine-tuning each category (ISC_*class*). Nevertheless, it is clear that if all the models were all fine-tuned better results would have been achieved. Within this principle, the framework presents low-complexity models for which is easy to specify the hyperparameters (all are the same between different ISC_*class*) and, possibly, have a faster training procedure than other more complex models presented in the state-of-the-art (see Section 3.2). In terms of training time, for the class ISC_ONMM, the train takes approximately 4 hours. In categories with fewer samples takes less time to train, and more samples more time, which varies from around 1 hour to around 7 hours (for the HISC model). The training of the ensembles is quite fast, taking a few minutes.

Finally, it is crucial to emphasize (as already mentioned) that tests and results are reported for each (individual) model/classifier - module in the overall framework (see Table 3.8 and Table 3.9) as well as for each ensemble classifier module. Nevertheless, no ablation study was conducted, as the focus was on the comparison of the models and the ensemble models. This means for instance, that the impact of employing combinations of two (individual) models for

the ensemble rather than the three models was not investigated (for each ensemble). In future work, we intend to conduct a full ablation study when utilizing increasingly complex individual modules.

3.6 CONCLUSIONS

This work presented research on text and image sentiment classification, especially in social media posts related to “indoor”, “man-made outdoors”, “non-man-made outdoors” and “indoor/outdoor with persons” environments. The study and analysis demonstrated the effectiveness of the proposed class-specific and holistic image classifiers and text classifiers in predicting sentiments, highlighting their potential applications and implications.

The pre-processing techniques, the incorporation of deep learning models and the advanced feature extraction techniques used led the Multimodal Sentiment Classifier framework to obtain high accuracy in sentiment classification from image-text posts, as well as a very consistent prediction of image and text represent the same sentiment and indeterminate sentiments.

Experiments on detecting sentiments in images showed promising results, demonstrating that the system can classify sentiments based on objects, colors and other aspects present in an image. Additionally, scene-specific image models obtained higher accuracies due to their ability to capture specific details in contexts, while the holistic model offers lower accuracy but higher versatility, once it is not needed to use pre-classification techniques to classify the images (the last being class segmentation, which is out of the focus of this chapter).

Finally, the ensemble models allowed the system to leverage the complementary information provided by the textual and visual models, which leads to better performance. This is very significant when a multiple model approach is used, in this case in sentiment analysis tasks.

In summary, this study contributed to the understanding of sentiment detection from data present in text and images. Although the accuracy of the system can be improved, the potential of the models has been demonstrated. More significantly, it introduced a framework that has not yet been published in the literature. The framework utilizes separate sentiment models for text and image; these models are only combined at the end if they convey the same sentiment or if there are uncertainties about the sentiment, allowing one to enhance the other. In the cases, where the image and language clearly convey different sentiments, they should not be

merged. An empirical conclusion is that the user only intended to illustrate the text or vice versa, that is, the text was only used to frame the image. This paper presents an initial approach to the discussion of this problem, that in future works need to be deepen. Continuing to improve and refine benchmark sentiment classification can open new possibilities to enable more sophisticated and nuanced sentiment analysis in interfaces and/or robots.

Looking ahead, there are several directions for future research. The focus should be on improving the sentiment detection model in images and obtaining more and better image sentiment classification datasets. We also intend to train and test models for images of scenes not mentioned during this research, to further increase the effectiveness of the final model that contains the responses from each environment-specific classifier.

4

CONCLUSIONS AND FUTURE WORK

This dissertation presented a research on text and image sentiment classification, especially in social media posts related to “indoor”, “man-made outdoors”, “non-man-made outdoors”, and “indoor/outdoor with persons” environments. The study and analysis demonstrated the effectiveness in predicting sentiments of the proposed category-specific and holistic image classifiers and text classifiers, highlighting their potential applications and strength.

The pre-processing techniques, the incorporation of deep learning models, and the advanced feature extraction techniques used, led the Multimodal Sentiment Classifier framework to obtain high accuracy in sentiment classification from text-image posts. It also ensured consistent predictions regarding whether the image and text convey the same sentiment or exhibit indeterminate sentiments.

Experiments on detecting sentiments in images showed promising results, demonstrating that the system can classify sentiments based on objects, colors, curves, and other aspects present in an image. Additionally, scene-specific image models obtained higher accuracies due to their ability to capture specific details in contexts, while the holistic model offers lower accuracy but higher versatility, once it is not needed to use pre-classification techniques to classify the images (being the last, category segmentation, out the focus of this chapter).

The ensemble models allowed the system to leverage the complementary information provided by the textual and visual models, which led to better performance. This is very significant when a multiple model approach is used, in this case in sentiment analysis tasks.

The main contributions of the dissertation are:

(i) the development of models to classify sentiments in natural scene images (landscapes);

(ii) the development of models to classify the sentiments linked with the text associated with a landscape;

(iii) the development of models to combine image and text classification, returning the information about its discrepancy;

(iv) a single hyperparameter image classification sentiment model that works in different scenes/environments (ONMM, OMM, IND, and IOwPB);

(v) a framework that combines image and text sentiment classification, returning the multimodal sentiment classification attach with the discrepancy (text-image sentiment) metric.

Looking forward, there are several directions for future research. The focus should be on improving the sentiment detection model in images and obtaining more and better image sentiment classification datasets. We also intend to train and test models for images of scenes not mentioned during this research, to further increase the effectiveness of the final model that contains the responses from each environment-specific classifier.

Another context of application would be to use the model to analyse the sentiment of videos and other multimedia content. This would require the development of a model that can analyse the sentiment of a sequence of images, which is a more complex task than analysing a single image. In a first approach, this would be a natural extension of the work presented in this dissertation, as the model could be used to analyse the sentiment of each frame of a video and then combine the results to obtain the sentiment of the video. In this context, other data could be added to the models, such as speech.

In summary, this study contributed to the understanding of sentiment detection from data present in text and images. Although the accuracy of the system can be improved, the potential of the models has been demonstrated. Continuing to improve and refine benchmark sentiment classification can open new possibilities to enable more sophisticated and nuanced sentiment analysis in interfaces and/or robots.

4.1 PUBLICATIONS

As already mentioned in Chapter 1, two scientific articles were developed, namely:

a. *Silva, N., Cardoso, P. J. S., & Rodrigues, J. M. F. (2024). Sentiment Classification Model for Landscapes,* In: Antona, M., Stephanidis, C. (eds), Universal Access in Human-

Computer Interaction. HCII 2024. Lecture Notes in Computer Science, vol. xx Springer, Cham. DOI: xx (waiting for the final proceedings).

b. *Silva, N., Cardoso, P. J. S., & Rodrigues, J. M. F. (2024). Multimodal Sentiment Classifier Framework for Different Scene Contexts*. Appl. Sci. 2024, 14, 7065. DOI: https://doi.org/10.3390/app14167065

REFERENCES

1. Mazali, Tatiana; From industry 4.0 to society 4.0, there and back. Springer, AI & SOCIETY, 2018, 7. DOI: 0.1007/s00146-017-0792-6
2. Kasinathan, P.; Pugazhendhi, R.; Elavarasan, R.M.; Ramachandaramurthy, V.K.; Ramanathan, V.; Subramanian, S.; Kumar, S.; Nandhagopal, K.; Raghavan, R.R.V.; Rangasamy, S.; et al. Realization of Sustainable Development Goals with Disruptive Technologies by Integrating Industry 5.0, Society 5.0, Smart Cities and Villages. Sustainability 2022, 14, 15258. DOI: 10.3390/su142215258.
3. Meena, G.; Mohbey, K.; Indian, A.; Khan, M.; Kumar, S.; Identifying emotions from facial expressions using a deep convolutional neural network-based approach. Multimedia Tools and Applications (2024) 83:15711–15732. DOI: 10.1007/s11042-023-16174-3
4. Kalla, D.; Smith, N.; Samaah, F.; Polimetla, K.; Facial Emotion and Sentiment Detection using Convolutional Neural Network. INDJAIR, Volume 1, Issue 1, January-December 2021, pp. 1–13. Available: https://iaeme.com/Home/issue/INDJAIR?Volume=1&Issue=1
5. Shneiderman, B. Human-Centered AI; Oxford University Press: Oxford, UK, 2022.
6. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 2022, 83, 19-52. DOI: 10.1016/j.inffus.2022.03.009.
7. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey, IEEE Trans Affect Comput, 2020, vol. 3045, no. c, pp. 1–20, DOI: 10.1109/TAFFC.2020.2981446.
8. Ruan, S.; Zhang, K.; Wu, L.; Xu, T.; Liu, Q.; Chen, E. Color Enhanced Cross Correlation Net for Image Sentiment Analysis. IEEE Trans. Multimed. 2021, 26, 4097–4109. https://doi.org/10.1109/TMM.2021.3118208.

9. Zhang, Z.; Luo, P.; Loy, C. C.; Tang, X. From Facial Expression Recognition to Interpersonal Relation Prediction, *Int J Comput Vis*, 2018, vol. 126, no. 5, pp. 550–569. DOI: 10.1007/s11263-017-1055-1.
10. Ekman, P. Are there basic emotions?, Psychological Review, 1992, vol. 99, no 3, pp.550-553. DOI: 10.1037/0033-295X.99.3.550.
11. Noroozi, F.; Corneanu, C. A.; Kaminska, D.; Sapinski, T.; Escalera, S.; Anbarjafari, G. Survey on Emotional Body Gesture Recognition, IEEE Trans Affect Compututing, 2021, vol. 12, no. 2, pp. 505–523. DOI: 10.1109/TAFFC.2018.2874986.
12. Nandwani, P.; Verma, R. A Review on Sentiment Analysis and Emotion Detection from Text. In Social Network Analysis and Mining; Springer: Berlin/Heidelberg, Germany, 2021; Volume 11. https://doi.org/10.1007/s13278-021-00776-6.
13. Ortis, A.; Farinella, G.M.; Battiato, S. An Overview on Image Sentiment Analysis: Methods, Datasets and Current Challenges. In Proceedings of the 16th International Joint Conference on e-Business and Telecommunications, Prague, Czech Republic, 26–28 July 2019; SciTePress: Setúbal, Portugal, 2019; pp. 296–306. https://doi.org/10.5220/0007909602900300.
14. Fugate, J.M.B.; Franco, C.L. What Color is Your Anger? Assessing Col-or-Emotion Pairings in English Speakers. Front. Psychol. 2019, 10, 206. https://doi.org/10.3389/fpsyg.2019.00206.
15. Amencherla, M.; Varshney, L.R. Color-Based Visual Sentiment for Social Communication. In Proceedings of the 15th Canadian Workshop on Information Theory (CWIT), Quebec City, QC, Canada, 11–14 June 2017. https://doi.org/10.1109/CWIT.2017.7994829.
16. Peng, Y. F.; Chou, T. R. Automatic Color Palette Design Using Color Image and Sentiment Analysis, in Procs IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2019. DOI: 10.1109/ICCCBDA.2019.8725717.
17. Plutchik, R. Chapter 1—A General Psychoevolutionary Theory of Emotion. In Theories of Emotion; Plutchik, R., Kellerman, H., Eds.; Academic Press: Cambridge, MA, USA, 1980; pp. 3–33. https://doi.org/10.1016/B978-0-12-558701-3.50007-7.
18. Munezero, M.; Montero, C. S.; Sutinen, E.; Pajunen, J. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text," IEEE Trans Affect Comput, 2014, vol. 5, no. 2, pp. 101–111. DOI: 10.1109/TAFFC.2014.2317187.

19. Gaspar, A.; Alexandre, L. A. A multimodal approach to image sentiment analysis, in Procs Intelligent Data Engineering and Automated Learning–IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I 20, pp. 302–309.
20. Vadicamo, L.; Carrara, F.; Cimino, A.; Cresci, S.; Dell'Orletta, F.; Falchi, F.; Tesconi, M. Cross-media learning for image sentiment analysis in the wild. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 308–317. Available online: http://www.t4sa.it (accessed on 01 August 2024).
21. De Oliveira, W.B.; Dorini, L.B.; Minetto, R.; Silva, T.H. OutdoorSent: Sentiment Analysis of Urban Outdoor Images by Using Semantic and Deep Features. ACM Trans. Inf. Syst. 2020, 38, 23. https://doi.org/10.1145/3385186.
22. Chatzistavros, K.; Pistola, T.; Diplaris, S.; Ioannidis, K.; Vrochidis, S.; Kompatsiaris, I. Sentiment Analysis on 2D Images of Urban and Indoor Spaces Using Deep Learning Architectures. 2022. Available online: https://www.mturk.com/ (accessed on 01 August 2024).
23. Hassan, S.Z.; Ahmad, K.; Hicks, S.; Halvorsen, P.; Al-Fuqaha, A.; Conci, N.; Riegler, M. Visual sentiment analysis from disaster images in social media. Sensors 2022, 22, 3628.
24. Du, Y.; Liu, Y.; Peng, Z.; Jin, X. Gated Attention Fusion Network for Multimodal Sentiment Classification. Knowl.-Based Syst. 2022, 240, 108107. https://doi.org/10.1016/j.knosys.2021.108107.
25. M. Katsurai and S. 'Ichi Satoh, "Image Sentiment Analysis Using Latent Correlations Among Visual, Textual, and Sentiment Views." [Online]. Available: http://sentistrength.wlv.ac.uk/
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. https://doi.org/10.1109/CVPR.2017.243.
27. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258. https://doi.org/10.48550/arXiv.1610.02357

28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
29. Silva, N.; Cardoso, Pedro J.S.; Rodrigues, João M.F. Sentiment Classification Model for Landscapes, Accepted to 18th International Conference on Universal Access in Human-Computer Interaction, part of HCI International, 2024, Washington DC, USA.
30. Ramos, C.M.Q.; Cardoso, P.J.S.; Fernandes, H.C.L.; Rodrigues, J.M.F. A Decision-Support System to Analyse Customer Satisfaction Applied to a Tourism Transport Service, Multimodal Technologies and Interaction, 2023, 7(1), 5. DOI: 10.3390/mti7010005.
31. Cardoso, P.J.S.; Rodrigues, J.M.F.; Novais, R. Multimodal Emotion Classification Supported in the Aggregation of Pre-trained Classification Models. In Computational Science; Mikyška, J., de Mulatier, C., Paszynski, M., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M., Eds.; Springer: Cham, Switzerland, 2023; LNCS Volume 10477. https://doi.org/10.1007/978-3-031-36030-5_35.
32. Novais, R.; Cardoso, P.J.S.; Rodrigues, J.M.F. Emotion Classification from Speech by an Ensemble Strategy. In Proceedings of the 10th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion, Lisbon Portugal, 31 August–2 September 2022; Association for Computing Machinery: New York, NY, USA, 2023; pp. 85–90. https://doi.org/10.1145/3563137.3563170.
33. Das, R.; Singh, T.D. Image–Text Multimodal Sentiment Analysis Framework of Assamese News Articles Using Late Fusion. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 2023, 22, 161.
34. Chen, D.; Su, W.; Wu, P.; Hua, B. Joint multimodal sentiment analysis based on information relevance. Inf. Process. Manag. 2023, 60, 103193.
35. Yadav, A.; Vishwakarma, D.K. A deep multi-level attentive network for multimodal sentiment analysis. ACM Trans. Multimed. Comput. Commun. Appl. 2023, 19, 15.
36. Kumar, P.; Malik, S.; Raman, B.; Li, X. CMFeed: A Benchmark Dataset for Controllable Multimodal Feedback Synthesis. arXiv preprint arXiv:2402.07640, 2024
37. Miah, M.S.U.; Kabir, M.M.; Sarwar, T.B.; Safran, M.; Alfarhood, S.; Mridha, M.F. A multimodal approach to cross-lingual sentiment analysis with ensemble of

transformer and LLM. Sci. Rep. 2024, 14, 9603. https://doi.org/10.1038/s41598-024-60210-7.

38. Yang, H.; Zhao, Y.; Wu, Y.; Wang, S.; Zheng, T.; Zhang, H.; Che, W.; Qin, B. Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey. arXiv preprint arXiv:2406.08068, 2024.
39. Deng, Y.; Li, Y.; Xian, S.; Li, L.; Qiu, H. MuAL: Enhancing multimodal sentiment analysis with cross-modal attention and difference loss. Int. J. Multimed Info. Retr. 2024, 13, 31. https://doi.org/10.1007/s13735-024-00340-w.
40. Mao, R.; Liu, Q.; He, K.; Li, W. Cambria, E. The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. IEEE transactions on affective computing, 2022, 14(3), 1743-1753. DOI: 10.1109/TAFFC.2022.3204972.
41. Hu, H. A Vision-Language Pre-training model based on Cross Attention for Multimodal Aspect-based Sentiment Analysis. In Proceedings of the 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China, 19–21 April 2024; pp. 370–375. https://doi.org/10.1109/CVIDL62147.2024.10603872.
42. CrowdFlower. Image Sentiment Polarity. 2015. Available online: https://data.world/crowdflower/image-sentiment-polarity (accessed on 10 May 2024).
43. Rodrigues, J.M.F.; Cardoso, P.J.S. Body-Focused Expression Analysis: A Conceptual Framework. In: Antona, M., Stephanidis, C. (eds) Universal Access in Human-Computer Interaction. HCII 2023. Lecture Notes in Computer Science, vol 14021. Springer, Cham., 2023, DOI: 10.1007/978-3-031-35897-5_42