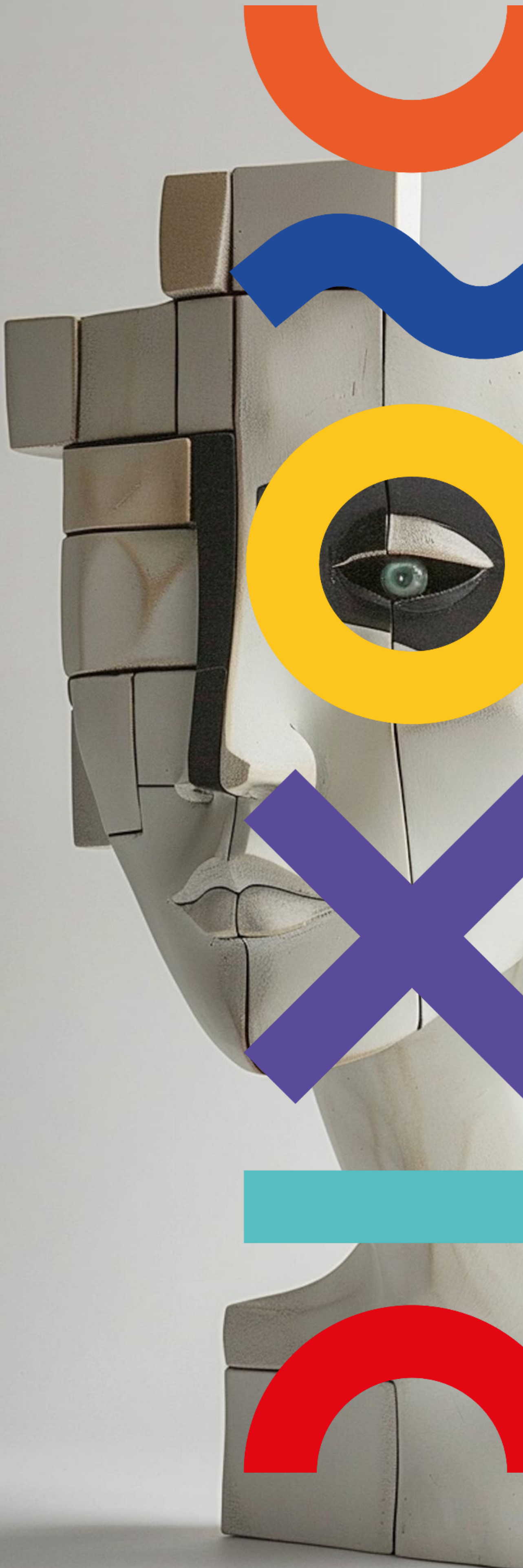
D3.1

Overview of the state of the art



D3.1 Overview of the state of the art

Dissemination Level: PU - Public
Lead Partner: PBY
Due date: 31.08.2024
Actual submission date: 31.08.2024

PUBLISHED IN THE FRAMEWORK OF

ENCODE - Unveiling emotional dimensions of politics to foster European democracy consumers

AUTHORS

Jim Ingbretsen Carlson, PBY
Rodrigo Ortega Izquierdo, PBY
Frans Folkvord, PBY

REVISION AND HISTORY CHART

VERSION	DATE	EDITORS		COMMENT
0.1	31/06/2024	Jim Ingebretsen Carlson	PBY	ToC
0.2	01/08/2024	Jim Ingebretsen Carlson, Rodrigo Ortega, and Frans Folkvord	PBY	First version of deliverable submitted for review
0.3	27/08/2024	Pawel Nowakowski	UWR	Reviewed
0.4	27/08/2024	Jim Ingebretsen Carlson	PBY	Comments addressed and sent for final review
1.0	31.08.2024	Aleksandra Oleksik	ASM	Submission to Participant Portal

DISCLAIMER

The information in this document is subject to change without notice. Company or product names mentioned in this document may be trademarks or registered trademarks of their respective companies.

All rights reserved

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein. Responsibility for the information and views expressed in the therein lies entirely with the author(s).

TABLE OF CONTENTS

LIST OF FIGURES:

LIST OF TABLES:

EXECUTIVE SUMMARY

Deliverable 3.1 (D3.1) – Overview of the state of the art, is the first deliverable within Work Package (WP) 3 - Analysing Social Media Communication. The overall aim of the WP is to analyse the interrelationship between emotions, values and identities in narratives by collecting a vast amount of social media discussions. The first step to achieve this is to identify and summarize the state of the art of emotion and sentiment detection techniques for text analysis, which will be done in D3.1. Since multiple up-to-date systematic reviews exist on the topic, the state of the art is generated by conducting a scoping umbrella review and complementary desk research covering the following areas: Psychological models of emotions employed in text analysis, pre-processing and feature extraction techniques and emotion detection techniques. The emotion detection techniques comprise the largest part of the evidence starting from lexicon-based algorithms and ending up at the state-of-the-art transformer models including Large Language Models.

1 INTRODUCTION

1.1 THE ENCODE PROJECT

The ENCODE project, titled "Unveiling Emotional Dimensions of Politics to Foster European Democracy," aims to explore and decode the role of emotions in political discourse and their impact on democratic processes. Recognizing that emotional appeals have significantly influenced political movements and voter behaviour, ENCODE seeks to understand the interplay between emotions, values, and identities. The project's primary goal is to create new positive narratives that can foster trust and engagement in European democratic processes, thereby counteracting the negative emotions that often dominate political discussions. Through innovative methodologies, including social media sentiment analysis, biometric research, and surveys, ENCODE aims to provide policymakers with tools and strategies to better incorporate the emotional needs of citizens into governance, ultimately enhancing democratic resilience and fostering a more inclusive political environment.

1.2 OBJECTIVES OF DELIVERABLE

The main objectives of WP3 - Analysing Social Media Communication, are related to transformation of communication using social media sites, and the concerns that are connected to political interference and the distribution of fake news and disinformation. We will analyse the interrelationship between emotions, values and identities in narratives by collecting a vast amount of social media discussions. Additionally, an analysis of emotions related to disinformation, in relation to social media output, will also provide significant results for understanding how European political and cultural spaces are evolving and influence the emotions in politics, along with an insight into some main problems of societal cohesion.

In this task, which is the first of WP3, we start by establishing the state-of-the-art by conducting a scoping umbrella review of the scientific evidence and a complementary desk research to provide the Consortium partners with the input for the methodological guidelines, along with an historical overview on social media analysis with specific focus on emotion and sentiment detection as well as Natural Language processing for analysing affects and emotions to provide a comprehensive overview of the state of the art of this methodology.

1.3 STRUCTURE OF THE DOCUMENT

The rest of the deliverable is structured in the following sections:

- Section 0 – Introduces the aim of the deliverable including the research questions and the methodology applied to answer them.
- Section 3 – Provides the results of the scoping umbrella review and complementary desk research, arriving at the state of the art of emotion and sentiment detection using text.
- Section 0 – Concludes the deliverable and outlines the next steps in WP3.

1.4 RELATION TO OTHER TASKS

This task and deliverable set the basis for all work conducted within WP3. Since it provides the state of the art of the methodology for conducting emotion and sentiment detection using text data, its output feeds straight into the work in all tasks of the WP3. Specifically, in

T3.2 - Build a methodological framework for social media analysis of emotions it provides the methodological part for how emotions will be detected in the social media analysis. Subsequently, it is the building block of T3.3 - Data collection and sentiment analysis of emotions in all countries, since the analysis is based on the emotions detected in the data. Finally, T3.4 - Establishing a catalogue of best practices aims at collecting the evidence of the influence of emotions on the political landscape and how to best incorporate this in democratic practices. Consequently, the evidence collected in this task is made possible thanks to the state of the art methodology outlined in this deliverable and subsequently applied in T3.2 and T3.3.

2 AIMS AND METHODOLOGY

The aim of the work in this task is to provide overview of the state-of-the-art of Natural-Language-Processing (NLP), Machine Learning (ML), Artificial Intelligence (AI) and Deep Learning (DL) techniques for conducting analysis of emotions and their applications to text data. To align with the main project goal - to decode the meanings of emotions and encode them to policy-making strategies aiming for a positive emotional turn - the main focus of this task will be on the analysis of emotions rather than on sentiments only. Sentiment analysis typically involves classifying text into whether its content is mostly negative, neutral or positive. Emotion analysis, as outlined in more detail below, provides more detailed information of the affects and emotions displayed in text. In fact, many models of emotions measure the sentiment at least indirectly by mapping the emotions on its degree of valence, which effectively is the degree of positivity/negativity of the emotion. Additionally, as the analysis conducted in WP3 will be done using textual data, the desk research will be limited to this mode of emotion expression.

To generate the state-of-the-art, we will conduct a literature review and complementary desk research. We have broken down the general aim of this task into three research questions that the review will answer. These are:

1. Which theoretical models of emotions are most used to detect emotions in text?
2. Which are the most common pre-processing and feature extraction methods of text that are used for emotion detection?
3. Which are the most common emotion detection techniques?

The first question covers the theoretical models and therefore sets the basis for what types of emotions that are typically analysed using text and social media data. A ML or AI model cannot understand raw text. Therefore, the second question aims to understand and outline the computational techniques for preparing text and transforming it into numerical representations that the emotion detection models can use to analyse text data. The aim of the third question is to provide an overview of emotion detection techniques that have been used in the scientific literature. The focus of this section will be on techniques to predict and detect emotions in text. Therefore, it comprises techniques such as lexicon- and rule based, traditional ML and deep learning. As the field has in general started with lexicon- and rule-based methods to ML and thereafter to deep learning using transformer architecture, the review provided here provides a historical overview of the techniques used to end up in the state-of-the-art methods.

At the start of the task, we assessed whether doing a systematic review on the topic was useful and necessary considering the existing scientific evidence. Therefore, the major scientific data bases PubMed, ScienceDirect, Scopus, Web of Science and google scholar were searched for (systematic) reviews on emotion and sentiment classification using social media data or text data in general. Specifically, the following search query was employed:

(emotion OR sentiment) AND (detection OR classification) AND ("machine learning" OR "deep learning" OR "artificial Intelligence" OR transformers) AND review

In doing so, a considerable amount of relevant systematic reviews was found that had been published in recent years. Therefore, it was decided to conduct a scoping umbrella review of the literature and present the comprehensive and extensive overview in this deliverable. An umbrella review is basically a review of systematic reviews and should be conducted when multiple updated systematic reviews exist on the topic of interest (Belbasi et al., 2022) which is our case. Conducting a scoping review suits our purpose well since it is recommended to be conducted "to examine how research is conducted on a certain topic or field" (Munn et al., 2018). Since this is exactly the aim of this task, conducting a scoping umbrella review and complementary desk research well generate a comprehensive overview of the state of the art of techniques for detecting emotions using text as an input will be achieved.

The literature review was done by the research team assessing whether the extracted scientific articles complied with the following inclusion criteria: 1) Dealing with techniques for emotion or sentiment detection, 2) Using ML or AI or DL technique 3) for text data 4) being some type of review (systematic, scoping etc) 5) Published in 2019 or later. 3) was not exclusive meaning that a review discussing techniques for text classification and other modes, such as videos or photos, would also be included. This generated a data set of relevant scientific reviews and Table 1 displays a selection of the most relevant systematic reviews on techniques for emotion and sentiment detection using text data that have been published since 2019. It shows the title, the main focus, the year of publication and the reference to each systematic review. Due to the vast use of social media data for emotion classification and sentiment extraction, articles of this type were typically always reviewed in the systematic reviews even though some did not limit their scope to social media data. Consequently, the numerous relevant systematic reviews on the topic displayed in Table 1 show that a scoping umbrella review is more adequate than a systematic review for generating the state-of-the-art.

Table 1. Selection of recently published systematic reviews covering techniques for emotion and sentiment detection.

Title	Main focus	Year	Reference
Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions	Machine learning techniques for emotion detection and sentiment analysis	2024	Alslaity & Orji (2024)
Using transformers for multimodal emotion recognition: Taxonomies and state of the art review.	Transformer architectures and their use for emotion recognition and classification from different modalities including text.	2024	Hazmoune & Bougamouza (2024).
Predicting emotions in online social networks: challenges and opportunities	Emotion detection in online social networks. Including models, computational techniques and data sets.	2022	Alqahtani & Alothaim (2022)
A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection	AI techniques for text-based emotion classification	2023	Kusal et al (2023)
A survey on deep learning for textual emotion analysis in social networks.	Deep learning techniques for analysing emotions in texts from social media	2022	Peng et al. (2022).
A Review of Different Approaches for Detecting Emotion from Text	Approaches, models, datasets, lexicons, metrics related to	2021	Murthy & Kumar (2021)

	emotion detection including in social media.		
A Survey of Textual Emotion Recognition and Its Challenges	Computational techniques and models for emotion recognition in text	2021	Deng & Ren (2021)
A survey of state-of-the-art approaches for emotion recognition in text	Computational techniques and models for emotion recognition in text	2020	Alswaidan & Menai (2020)
Sentiment analysis using deep learning techniques: a comprehensive review	Deep learning techniques for sentiment analysis	2023	Sahoo et al. (2022)
Transformer models for text-based emotion detection: a review of BERT-based approaches	Transformer models for detecting emotions in text.	2021	Acheampong et al. (2021).
Review on sentiment analysis for text classification techniques from 2010 to 2021	Techniques for sentiment analysis across different domains	2023	Ullah et al. (2023)
Textual emotion detection in health: Advances and applications	Emotion detection in text with focus on health and medicine	2023	Saffar et al. (2023)
Emotion detection in text: A review	Computational techniques and models for emotion recognition in text	2018	Seyeditabari et al. (2018)
A systematic review on affective computing: emotion models, databases, and recent advances	Affective computing, including emotion detection in text and other modalities	2022	Wang et al. (2022)
Sentiment analysis using deep learning architectures: a review.	Sentiment analysis and deep learning	2020	Yadav, & Vishwakarma (2020)
A review on sentiment analysis and emotion detection from text	Sentiment analysis and emotion detection from text. Models and techniques.	2021	Nandwani & Verma(2021)
Transformers in Machine Learning: Literature Review	Transformers for Natural Language Processing tasks including emotion detection	2023	Thoyyibah et al. (2023)
Sentiment analysis in social media and its application: Systematic literature review.	Techniques, models, data sets, topics for sentiment analysis using social media data	2019	Drus & Khalid (2019).
Sentiment analysis: a review and comparative analysis over social media.	Techniques and models for sentiment analysis using social media data	2020	Singh et al. (2020).
Deep learning and multilingual sentiment analysis on social media data: An overview.	Deep learning techniques for multilingual sentiment analysis using social media data	2021	Agüero-Torales et al. (2021).
A literature review on sentiment analysis techniques involving social media platforms.	Techniques for sentiment analysis using data from Twitter	2020	Garg et al. (2020)
Systematic Review of Emotion Detection with Computer Vision and Deep Learning	Deep learning techniques for emotion detection from facial and body expressions	2024	Pereira et al. (2024).

Source: Author's own elaboration

3 EMOTION DETECTION IN TEXT DATA

This chapter answers our three research questions by displaying the results of the scoping umbrella review and complementary desk research. As the ENCODE project and consequently this WP (WP3) is centred around the study of emotions, it starts out by outlining the most common models of emotions used in social media analysis in Section 1. These models typically stem from research in psychology and have been operationalised in various ways for emotion detection in text. Following this, Section 0 outlines the methods for pre-processing text data (Section 0) and for feature extraction (Section 1), which are two important steps in preparing the data when applying traditional ML models. Thereafter, the main body of results follow where we outline the techniques for emotion detection that have been used leading up to what is currently the state of the art in Section 3.3. This section is divided into four sections starting with the simpler Lexicon approaches outlined in 3.3.1. Thereafter, traditional ML used for emotion classification is presented in 3.3.2. Approaching the state of the art, Section 0 outlines the early DL models for emotion detection which are all variants of neural networks. We reach the state of the art in Section 2, where we provide an overview of the recent advances in deep learning which are centred around the models using the transformer architecture. To give a more in-depth understanding into how different DLmethods have been used in the literature, Section 3.3.5 provide some examples from the literature how this has been done and the results they achieved. Being examples, this section should not be understood as a comprehensive view of the literature.

3.1 EMOTION MODELS

The study of emotions has received a lot of attention in the scientific literature. In general, emotions refer to a basic nervous system related to a mental state, such as joy, anger, or sadness and it is also being as episodes affected by stimuli. More specifically, emotion was defined as “an episode of interrelated synchronized changes in the state of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism” (Scherer, 2005). However, more than 90 definitions of emotions have been offered and there exist almost an equal number of theories of emotion. Naturally, there are many categorisations of emotion such as cognitive versus non-cognitive, instinctual versus cognitive, and categorisations based on time, as some emotions have a duration of seconds whereas others can last for years (Cambria et. al., 2012).

In this deliverable, we narrow our focus to the models of emotions that have been used for emotion detection using text data. Błąd! Nie można odnaleźć źródła odwołania. shows an overview of some of the most common psychological models of emotion used.

Table 2. Common emotion models used for emotion detection using social media data

Model and author	Emotions	Type
Ekman (1992)	anger, disgust, fear, happiness, sadness, and surprise.	Categorical
Shaver (1987)	Anger, fear, joy, love, sadness, surprise	Categorical
Plutchik (1980)	Acceptance, admiration, aggressiveness, amazement, anger, annoyance, anticipation, apprehension, awe, boredom, contempt, disapproval, disgust, distraction, ecstasy, fear, grief, interest, joy, loathing, love, optimism, pensiveness, rage, remorse, sadness, serenity, submission, surprise, terror, trust, vigilance	Dimensional
OCC	Admiration, anger, appreciation, disappointment, disliking, fear, fears confirmed, gloating,	Dimensional

Ortony A, Clore GL & Collins A (1988)	gratification, gratitude, happy-for, hope, liking, pity, pride, sorry-for, relief, remorse, reproach, resentment, self-reproach, shame.	
Lövheim (2012)	Anger, disgust, distress, enjoyment, fear, interest, shame, surprise	Dimensional
Circumplex Russell (1980)	Anger/rage, contempt/disgust, distress/anguish, enjoyment/joy, fear/terror, interest/excitement, shame/humiliation, surprise/startle	Dimensional
Hourglass Cambria et al. (2012)	Ecstasy, vigilance, rage, admiration, joy, anticipation, anger trust, serenity, interest, annoyance, acceptance, pensiveness, distraction, apprehension, boredom, sadness, surprise, fear disgust, grief, amazement, terror, loathing	Dimensional

Source: Author's own elaboration

Models of emotions are often described as either categorical or dimensional. The categorical models describe emotions as discrete and distinct. The dimensional models, use a number of underlying dimensions to infer the emotions. Most models use the dimensions of valence (positive or negative), arousal (the degree of excitation), and dominance (level of control over the emotion). Consequently, an emotion has a certain degree of valence, arousal, and dominance. In contrast to the categorical models, the dimensional models therefore allow to measure the intensity of the different emotions based on the underlying dimensions. The hourglass model by Cambria et al. (2012) is designed to specifically capture human-computer interaction and how humans emotionally react to written text. It uses the following four underlying dimensions of emotions: Pleasantness, (amusement by interaction modalities), Attention (interest in interaction contents), Sensitivity (comfort with interaction dynamics), and Aptitude (confidence in interaction benefits). The systematic review by Alqahtani & Alothaim (2022) finds that the Ekman model of emotions is the most used model for emotion detection using social media data. However, their review suggests that many different models are frequently applied.

3.2 METHODS FOR PRE-PROCESSING AND FEATURE EXTRACTION

Text data need to be prepared for an AI model to be able to use it. First, it needs to be cleaned and prepared in various ways. These techniques for pre-processing are outlined in Section 0. The AI models cannot understand text. Therefore, the text needs to be transformed into numerical representations in one way of another. The most common techniques for achieving this are collected in Section 1.

3.2.1 PRE-PROCESSING

Pre-processing is a fundamental step in NLP that prepares raw text for analysis and ensures that the data is clean, consistent, and suitable for ML models. It encompasses several critical tasks including data cleaning, tokenization, and normalization, each playing a vital role in transforming text into a format that algorithms can effectively process and analyse. By meticulously performing these preprocessing steps, we can enhance the quality of the textual data and improve the accuracy and performance of NLP applications (Singh et al., 2020; Kusal et al., 2023). Additionally, pre-processing, as any other process dealing with data provided below, is carried out in a statistical software such as Python, R, or STATA. It is carried out by writing code which is easily shared among researchers and data scientists. This allows for replicability of the results and that the process is transparent.

Data cleaning in NLP involves several steps to prepare raw text for analysis. This includes converting all text to lowercase to maintain uniformity, removing punctuation and digits, eliminating stop words that do not contribute significant meaning, and stripping out hashtags and HTML tags. Additionally, expanding contractions to their full forms ensures consistency throughout the text. These cleaning steps are essential to remove noise and irrelevant information, making the text more suitable for subsequent analysis.

Tokenization, or segmentation, is the process of splitting text into smaller units called tokens, which can be words, sub-words, or characters. This step is crucial for most NLP tasks as it breaks down the text into manageable pieces that algorithms can process. Word tokenization splits sentences into individual words, sub-word tokenization handles rare words and different word forms, and character tokenization divides text into individual characters. Additionally, sentence tokenization separates text into individual sentence. Proper tokenization is fundamental to accurately analyse and interpret textual data since to understand large texts, the machine needs to understand the highly complex relationships between the smaller pieces of texts of which they are composed. These pieces are the tokens.

Normalization involves reducing words to their root or base form, a process that includes stemming and lemmatization. Stemming truncates words to their base form, often by removing suffixes, without necessarily producing a real word. Lemmatization, on the other hand, converts words to their base or dictionary form, considering the context and parts of speech to ensure accuracy. This step reduces the variability in the text and helps in recognizing different forms of the same word as a single entity. Normalization is vital for simplifying text data and improving the performance of NLP models by ensuring consistency in how words are represented.

3.2.2. FEATURE EXTRACTION (FOR TEXT CLASSIFICATION)

Feature extraction is a crucial process in data analysis and machine learning that involves transforming raw(text) data into a set of characteristics or features that can be effectively used by predictive models. This process is essential in simplifying the data, reducing its dimensionality, and making it suitable for analysis while retaining the most significant information.

Feature extraction techniques can be broadly classified into traditional approaches and modern approaches based on machine learning and deep learning. Traditional approaches focus on syntactic representations of data, such as Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), Part of Speech (POS) tagging, Named Entity Recognition (NER), and Feature Hasting (FH). These methods are effective in representing text data by converting it into numerical vectors, which can be processed by machine learning algorithms (Kusal et al., 2023; Singh et al., 2020; Eskandar, 2023; Saffar et al., 2023; Ullah et al., 2023).

Traditional Approaches

- BoW counts the frequency of words in a text and generates a sparse vector that represents the presence or absence of words. It is often used in text classification tasks where the frequency of each word is considered as a feature. Common variations include n-grams, which capture the frequency of contiguous sequences of words.
- TF-IDF evaluates the importance of a word in a document relative to its frequency across a collection of documents. It helps in identifying significant words within specific documents, thereby enhancing the effectiveness of the features used in text analysis.
- POS tagger identifies the grammatical categories of words, such as nouns, verbs, adjectives, and adverbs. It provides the lowest level of syntactic analysis for parsing and word sense disambiguation. This information is useful for understanding the structure and meaning of sentences, which can be critical in various NLP tasks.

- NER involves identifying and classifying entities in text into predefined categories such as names of people, organisations, locations, dates, etc. This helps in extracting structured information from unstructured text.
- FH tagged words are labels of sentiments and emotions embedded by the writer itself. Hashtags are extremely useful for extracting emotion from social media data.

Modern approach

- Word embeddings represent each word as a vector in a high-dimensional space where similar words are close together and dissimilar words are far apart. Word embeddings are typically learned from large amounts of text data using unsupervised learning methods like Word2Vec and GloVe. The basic idea is to use a neural network to predict the context words of each target word in a large text corpus. The learned weights of the neural network are used as the word embeddings. Once the word embeddings are learned, they can be used to represent words in downstream NLP tasks like sentiment analysis and text classification (for emotions) (Seyeditabari et al.,2018; Deng & Ren, 2021).

3.3. EMOTION DETECTION TECHNIQUES

This section outlines the main body of results from the scoping umbrella results which concerns the techniques for detecting and predicting emotions in text. It starts out in Section 3.3.1 with the more basic and less data-driven lexicon approaches that are based on keywords or linguistic rules to detect emotions. Subsequently, we outline the traditional ML models used for emotion detection in Section 3.3.2. We approach the state-of-the-art in the next section that summarize the early DL classifiers in Section 3.3.3. Thereafter, Section 2 outlines the latest developments and collects the models using the transformer architecture with the LLMs as a special case. Since the previously mentioned chapters describe the most common technique used, the section ends with a section with examples of how DL techniques have been applied for emotion detection. This is presented in Section 3.3.5.

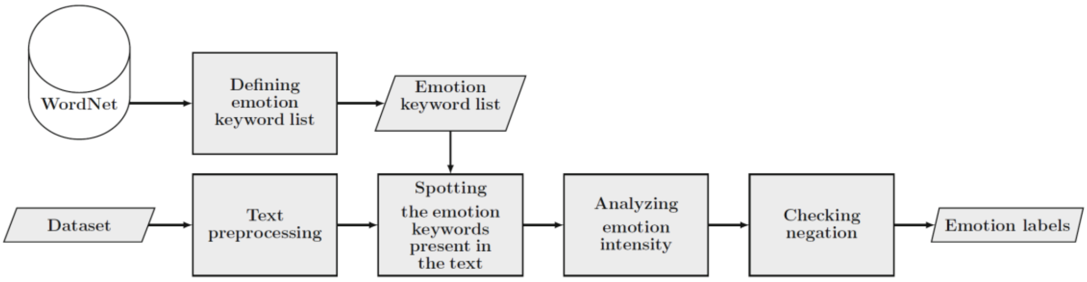
3.3.1. LEXICON APPROACHES

Lexicon approaches are techniques that rely on pre-prepared rules or scores stemming from theoretical or empirical linguistic studies that characterize the meanings and syntactical structures of words and sentences. Usually, these methods are applied to sentiment analysis given the positive or negative nature of a given word, wording or sentence, but it can be expanded to help in emotion classification. However, the higher complexity of such task imposes certain limitations to lexicon-based approaches in textual emotion classification (Singh et al., 2020; Drus & Khalid, 2019; Wang et al. 2022). Lexicon approaches can be divided into Keyword-based, explained in Section 1 and rule-based that are outlined in Section 3.3.1.2.

3.3.1.1. KEYWORD-BASED

The keyword-based techniques identify keyword occurrences in text and compares them to the annotations in a reference data set. In this process, keywords that relate to each of the studied emotions are derived from conventional lexical resources such as WordNet or WordNet-Affect, or specifically created for the task. Consequently, certain words, or keywords, are related to diverse types of emotions. After pre-processing the dataset, keyword matching is conducted between the predefined keyword list and the emotion words in the text. The intensity of the emotion keywords is then examined, followed by an analysis of negation to determine its presence and scope. Finally, the emotion tag calculation is performed based on the presence of keyword their relative weight and their associated emotion weights, as well as the presence of negation which could indicate a negative emotion (Kusal et al., 2023; Saffar et al., 2023).

Figure 1 Main steps of a keyword-spotting technique.

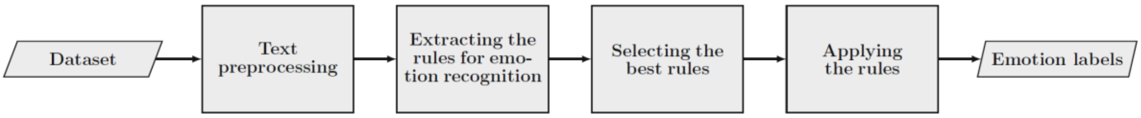

WordNet
Defining emotion keyword list
Emotion keyword list
Dataset
Text preprocessing
Spotting the emotion keywords present in the text
Analyzing emotion intensity
Checking negation
Emotion labels

Source: Alswaidan & Menai (2020)

3.3.1.2. RULE-BASED

Rule-based approaches utilize statistical linguistic rules, meaning associations between a particular linguistic pattern and a tag, to recognize emotions from the text. These rules are built using statistical and linguistic concepts, with probabilistic affinity stored for each word. This set of rules or lexicon identifies a word's sentiment or opinion and their corresponding intensity measure. First, text pre-processing is performed on the emotion dataset. The pre-processing steps may include tokenization, stop words removal, lemmatization, POS tagging, and dependency parsing. Then, the emotion rules are extracted using linguistic, statistics, and computational concepts. The best rules are then selected. Finally, the rules are applied to the emotion dataset to determine the emotion labels. (Alqahtani & Alothaim, 2022; Alswaidan & Menai, 2020; Kusal et al., 2023; Saffar et al., 2023). Udochunkwu & He (2015) use the rule-based method for emotion detection for text by identifying subject-verb relationships and assessing the polarity of actions and events using lexicon-matching techniques. It applies pre-defined rules to classify emotions based on contextual factors, allowing it to effectively detect implicit emotions without relying on explicit emotion-bearing words, and has shown significant performance improvements over traditional methods in various datasets.

Figure 2 Main steps of a rule-based approach.



Source: Alswaidan & Menai (2020)

3.3.2. TRADITIONAL MACHINE LEARNING CLASSIFIERS

Lexicon-based methods for text classification, particularly in the domain of emotion detection, rely heavily on human-driven processes, where keywords and rules in the lexicon are derived from linguistic theoretical frameworks or empirical textual analysis. These traditional methods have been valuable for their interpretability and alignment with established linguistic theories. However, the emergence of ML techniques introduces a new horizon of possibilities, enabling more complex and automated approaches to emotional classification of text. ML techniques can be categorized into supervised and unsupervised methods, both offering unique advantages for text classification tasks. In supervised learning, models such as decision trees or support vector machines (SVM), are trained on a labelled dataset, where each text instance is manually coded with the correct emotion labels. This training set serves as the foundational experience for the model, allowing it to learn patterns and features (typically words) associated with different emotional categories. The model's performance is evaluated using a test set—an unseen portion of data not used during training—by measuring how many emotions it correctly predicts compared to the manually coded labels in the test set, determining its effectiveness and generalizability. On the other hand, unsupervised learning does not rely on labelled data but seeks to identify intrinsic patterns and structures within the text through techniques like clustering (e.g., K-means, hierarchical clustering) and topic modelling (e.g., Latent Dirichlet Allocation). These clusters or topics can be analysed to infer potential emotional categories (Garg et al., 2020; Alslaity & Orji (2024); Ullah et al., 2023).

Machine learning models offer significant advantages such as scalability, adaptability, and automation. They can process vast amounts of text data quickly, adapt to new data, and evolve over time, and classify emotions in text automatically, reducing the workload for analysts and researchers. However, challenges such as ensuring high-quality and representative training data, maintaining interpretability, and addressing bias and fairness concerns must be considered. Integrating machine learning techniques, including both traditional and deep learning methods, into the emotional classification of text enhances the ability to analyse and understand emotions, providing valuable insights and practical applications across various domains. Table 2 displays the most common ML models for emotion detection using text. For each model, it contains a description of how it works and references to scientific papers applying the model for emotion detection.

Table 3 Most common algorithm for supervised ML emotion text classification.

Model	Description	Examples used for emotion detection
Naïve Bayes	Naive Bayes is a machine learning classifier and it used to solve classification problems. It uses Bayes theorem extensively for training. It can solve diagnostic and predictive problems. Bayesian Classification provides a useful point of view for evaluating and understanding many learning algorithms. It calculates explicit probabilities for hypothesis, and it is robust to noise in input information. In this multilabel classification, single Naive Bayes model is trained for predicting each output variable.	(Sendari et al., 2020; Parvin & Hoque, 2021)
Support Vector Machine	The support vector machine is a supervised learning distance-based model. It is extensively used for classification and regression. The main aim of SVM is to find an optimal separating hyperplane that correctly classifies data points and separates the points of two classes as far as possible, by minimizing the risk of misclassifying the unseen test samples and training samples. It means that two classes have maximum distance from the separating hyperplane.	(Parvin & Hoque, 2021; Zhao, 2021)
Random Forest	It is an ensemble learning method for classification and regression. Each tree is grown with a random parameter and the final output is achieved by aggregating over the ensemble. It is a classifier that contains a number of decision trees on different subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Rather than depending on one decision tree, the random forest takes the prediction from each tree and based on the	(Vora et al., 2017; Parvin & Hoque, 2021)

	majority votes of predictions, and it predicts the final output.	
K-Nearest Neighbour (KNN)	K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new data and available data and put the new data into the category that is the most similar to the available categories. It reserves all the available data and classifies a new data point based on the similarity. This means when new data comes out then it can be easily classified into a well suite category by using KNN algorithm. It can be used for Classification as well as for Regression but mostly it is used for the Classification problems. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a different category that is much similar to the new data.	(Parvin & Hoque, 2021; Zhao, 2021)

Source: (Kher & Passi, 2022)

3.3.3. DEEP LEARNING CLASSIFIERS

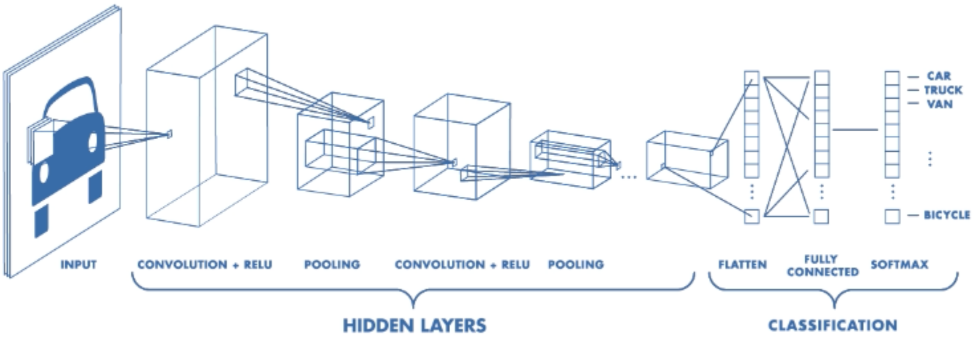
Within the broader field of ML lies DL, a subset that utilizes neural networks with many layers (hence "deep") to model complex patterns in data. DL models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have shown remarkable performance in understanding context and sequential data in text. These models are capable of automatically learning feature representations from raw data, which contrasts with traditional ML models that often require manual feature engineering. DL models can capture intricate patterns and dependencies in large datasets, making them highly effective for tasks like emotion classification, where understanding nuanced context is crucial. Lately, there has been great progress within this filled with the development of transformer models and notably the Large Language Models (LLMs). In this section, we divide the deep learning into the three categories: CNN, RNN, and LSTM. The subsequent section outlines the state of the art of DL including transformers and LLMs as a special case (Wang et al., 2022; Acheampong et al., 2021; Yadav, & Vishwakarma, 2020; Pereira et al., 2024)

Starting with the basics, an Artificial Neural Network (ANN) consists of several components: 1) An input layer 2) a number of hidden layers 3) an output layer. Each layer consists of a number of nodes (neurons) that are connected across layers with edges (Kusal et al., 2023; Hazmoune & Bougamouza, 2024). The input layer feeds the model with the features, in our case typically some numerical representation of words. Thereafter, these are passed onto the hidden layers as defined by the structure of the ANN (the edges and nodes). Weights are estimated for each edge to assess the strength of the connection between different nodes in the networks. At each hidden node, the input is multiplied by its weight and passed through an activation function, which allows to capture non-linearities in the data. Finally, the resulting values of the hidden layers are passed to the output layer which provides the prediction. In the case of classification, as of emotions, the values in the output layers are passed through a function converting them into probabilities of the emotions being present in the text. Throughout the training of the ANN, the weights are adjusted to minimize the error of the model i.e. to ensure that the model gives the most accurate prediction as possible. In our case, this means that the weights are adjusted to provide the most accurate prediction of the emotions present in the text.

3.3.3.1. CNN

One of the most popular families of DL methods are the CNNs. They are specifically designed to address classification and computer vision tasks and are widely use in image classification. Their structure is formed by a set of convolutional and pooling layers, that are then communicated with the fully connected layer that acts as classifier and produces the output. Figure 3 shows a graphical representation of how the CNN works with a hidden layer containing the convolution and pooling layers, and then a classification layer, using a neural network structure. As can be seen in the picture, the CNNs work such that the hidden layers filter the huge amount of input data into smaller pieces only containing the most relevant information (data). Subsequently, the classification model uses this more manageable and relevant data set as input.

Figure 3 Architecture of a CNN.



Source: MATLAB - What Are Convolutional Neural Networks? | Introduction to Deep Learning[1]

First let's picture an image like a matrix composed of pixels, the matrix will have NxM dimensions if it is in a grey scale or NxMx3 if it uses the RGB colour scheme. When an image is inputted into the CNN the first convolution layer applies a set of filters to the image. These filters are smaller matrices, with different weights as each of their values, and they multiply the input image moving at a given stride, generating what is called an activation map. To put this simpler, imaging one of these filters being a 3x3 matrix and with value 1 in the middle column describing a straight line, and a value 0 in the outer/lateral columns. The filter will just be moving one pixel at a time to the right, generating an activation map that will reinforce the shape the filter has, in this case the straight line. This process is repeated along different segments of the image in a systematic way, as so the first layer of the CNN will be able to identify basic shapes. After each convolutional layer we find a pooling layer that replaces the output of the network at certain locations by deriving a summary statistic of the nearby outputs. This helps in reducing the spatial size of the representation, which decreases the required amount of computation and weights. Basically, it takes the higher value in each 3x3 portion of the matrix, producing a new input to the next layer. As we go deeper into the CNN these features extracted by the convolutional layer will increase in complexity. Finally, these features are passed into the fully connected layer, that following a same process described earlier for ANN is able to classify that image given the features it is transmitted from the convolutional and polling layers process (Stankovic & Mandic, 2022; Kusal et al., 2023; Yadav, & Vishwakarma, 2020; Deng & Ren, 2021; Agüero-Torales et al., 2021).

[1] https://www.mathworks.com/discovery/convolutional-neural-network.html

3.3.3.2. RNN

RNNs were developed to better capture relationships in time-series data or when the relationships in the order of the input and output data matters. RNNs have been successfully used for various NLP tasks, such as emotion classification, because the order of the words in a sentence matter for the reader's (being a machine or a human) ability to understand it. While the ANN estimates the importance of each word independently of the other words in a sentence, the RNN estimates the importance of each word also taking into account the words previously appearing in the sentence.

This is achieved by ordering the inputs sequentially to create different time steps, in our case from the first word to the last word of an input text. For each time step, a hidden state, or memory state, is calculated. The hidden state captures the knowledge the network has up until this time step. In our case, the knowledge is the words appearing before the word at the current time step, their order, and their importance for making an accurate prediction. Different from an ANN, at each time step a prediction is carried out, for example of the next word appearing in the text given the word and knowledge (hidden state) up until that point. To improve these predictions, weights are given to both the input (word) and the hidden state (previous words and their importance) to decide their importance for an accurate prediction at each time step. The weights are then optimized through training by minimizing the prediction error, taking into account the error of all time steps. So, essentially, an RNN comprises a vast number of neural networks that are connected sequentially and that feed information forward through the hidden states until a final prediction is achieved. (Peng et al., 2022; Murthy & Kumar, 2021)

3.3.3.3. LSTM

While powerful for various NLP tasks, the RNNs suffer the problem that they have difficulties learning to store knowledge for a long time. In our case this is problematic since this means that an RNN may have problems making accurate predictions when knowledge appearing early in a long text is important for predicting the next word at the end of the text. Therefore, Hochreiter and Schmidhuber (1997) introduced Long Short Term Memory (LSTM) to deal with such long-term dependencies of different pieces of text to understand its content. The LSTM has a similar recurrent structure as the RNN, with information passed through each time step. To achieve long memory, cell states is introduced that runs through all time steps in the LSTM. Each cell state only has a minor linear interaction, in which relevant information can change, at each time step. This makes it easy for the cell states to carry information for a long time. However, at each time step the information in the cell state can be modified. First, a "forget gate" tells the cell states which information (words) from the previous time steps that are not relevant anymore and thus should be forgotten. Thereafter, in an "input gate", relevant new information (if any) is added to the cell state based on the new information (the word inputted and the hidden state at the current time step) (Sahoo et al., 2022; Nandwani & Verma, 2021; Thoyyibah et al., 2023; Agüero-Torales et al., 2021)

Many updates and variations exist of the LSTM. For example, the Bidirectional Long Short Term Memory (Bi-LSTM) captures the fact that in addition to knowing the preceding words, it can also be valuable to know the succeeding words to understand a sentence. To achieve this the Bi-LSTM combines two LSTMs: One as described above where the sequence of words inputted start at the first and finish at the last, and a second one which the sequence of words starts at the last word and finish at the first. Both these LSTMs provide a probability for the prediction, such as which emotion is expressed in the text, and their probabilities are combined to provide the final prediction. Additionally, the Gated Recurrent Unit (GRU), introduced by Cho, et al. (2014), is another well-used model that combines the forget gate and input gate into one "update gate".

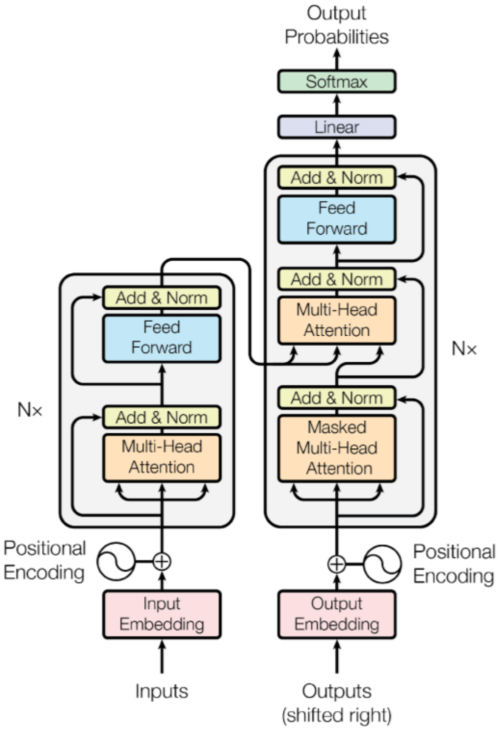
3.3.4. STATE OF THE ART DEEP LEARNING MODELS – TRANSFORMERS

One of the challenges that RNNs and CNNs faced are the lack of long-term memory, understanding long-term memory as the capacity to contextualise words beyond their immediate predecessors. To overcome this challenge, transformer-based deep learning models uses a mechanism called self-attention to weigh the importance of different parts of the input data. This concept of attention lies at the core to understand the power of Transformers in text generation. It was developed by Vaswani et al. (2023) under the title “Attention Is All You Need”. They designed an attention mechanism that computes a score for each word in the input sequence, determining how much focus each word should receive when producing a word in the output sequence. This model has proven to be superior in quality for tasks like language translation, as evidenced by its performance on benchmark datasets for English-to-German and English-to-French translation tasks, achieving state-of-the-art results. In a high-level approach, an attention layer updates the embeddings of each of the tokens moving its direction in the space to accommodate its meaning by inputting information from other preceding embeddings. In a much-simplified manner there is a set of keys and queries that activate relationships among words complementing their meaning given the context in which they are found.

However, attention is just an element in the whole transformer architecture, which is based on an encoder-decoder scheme (see Figure 4). This architecture consists of an encoder to process the input and a decoder to generate the output, both built using layers of multi-head self-attention and feed-forward neural networks. This architecture is used for tasks where an input sequence needs to be transformed into an output sequence, like translating a sentence from one language to another, generating text or classifying it. The encoder processes the input sequence and transforms it into a fixed-size representation called a context vector. This vector captures the essential information of the input sequence. The decoder then takes this context vector and generates the output sequence one element at a time, updating its internal state as it goes along.

The attention mechanism enhances the process by allowing the decoder to focus on different parts of the input sequence for each step of the output generation. Instead of relying on a single context vector, the attention mechanism dynamically computes the relevance of each input element when producing each output element. This results in a more flexible and accurate transformation, as the model can "attend" to the most relevant parts of the input for each part of the output.

Figure 4 The Transformer - model architecture.



Source: Vaswani et al. (2023)

These models have been widely used in recent year, and it has been shown that scale matter, as the number of parameters and training data are key in defining the performance of the model (Casola et al., 2022). This leads to a fundamental problem of training capacity both in available data and on computational power. Therefore, we observe the emergence of transformer pre-trained models. These models are transformers that have already been trained and can perform generic tasks of text generation on a standard manner, while allowing for further fine-tuning for more specific tasks such as emotion classification at much smaller cost on training capacity and training information. This makes it more practical to build on these pre-trained models as the whole contextual text understanding process has already been learned in the pre-training phase. Consequently, it allows to use only a relative well curated small sample to "teach" them a specific task, as this will only imply minor fluctuations of the already established parameters. To have a better perspective of such pre-trained models we summarize the most well-known and used models from the early adoption to the present state-of-the-art in the following sections.

3.3.4.1. BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

BERT's key technical innovation is applying the bidirectional training of Transformer attention model to language modelling. As opposed to directional models, which read the text input sequentially (from left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word). Devlin et al. (2019) show that a language model which is bidirectionally trained can have a deeper sense of language context and flow than single-direction language models (Horev, 2018).

BERT can understand language by training on the Masked Language Modeling (MLM) and the Next Sentence Prediction (NSP) mechanisms, which allows for this bidirectional training. It takes in as input some random sentences, masks some of the words in the sentences, essentially hiding the words, and then reconstructs the masked words from the surrounding texts at the output. The reconstruction is made by predicting the most likely word based on the surrounding words. Its ability to input two sentences at once and determine if the second sentence comes after the first makes it achieve NSP. This ability helps the model to maintain long-distance relationships between texts (Acheampong et al., 2021).

3.3.4.2. XLNET - GENERALIZED AUTO-REGRESSIVE MODEL FOR NLU

XLNet is based on the BERT architecture. However, it introduces a significant change in its training as it uses the permutation objective. Using Permutation Language Model (PLM), the model can learn bi-directional context by training all possible permutations of words in a sentence. Then using the positional encoding and recurrence mechanisms in Transformer-XL eradicates the fixed-length problem in BERT (Yang et al., 2020; Alswaidan & Menai, 2020).

3.3.4.3. GENERATIVE PRE-TRAINED TRANSFORMER (GPT)

GPTs leverages the semi-supervised learning approach to model language using transformer decoders (Radford et al. 2018). Mainly used for text representation, the GPT is made up of 12 transformer layers and 12 attention heads. The transformer has a decoder that uses the massive unlabelled datasets to understand language through pre-training and then fine-tunes on the limited supervised datasets to, for example, act as a chat bot and answer people's questions. The sole task of the GPT is to predict the next token in the sequence. The GPT's input is the input texts' weight embeddings plus their positional embeddings for context extraction. The input is passed to the multi-head attention layer in the 12 layered transformer decoder blocks, a feed-forward layer, and then the softmax outputs a probability distribution. Successive versions of GPT relay on bigger training corpus in order to improve the coherence of the output text (Acheampong et al., 2021).

3.3.4.4. LARGE LANGUAGE MODEL META AI (LLAMA)

LlaMa builds on the transformer architecture with several key enhancements. It introduces grouped multi-query attention, which speeds up inference by allowing different query heads to share the same keys and values, addressing memory access bottlenecks. Instead of the standard layer normalization, LlaMa uses Root Mean Square (RMS) normalization to regularize summed inputs by re-scaling them, enhancing stability and performance. Llama also employs rotary positional embeddings, which dynamically learn positional representations during training, unlike the static positional encoding vectors used in the original transformer. The token representations vary with model size, significantly increasing the dimensionality compared to the original transformer, and its input embeddings are learned during training. The model uses a key-value cache during inference to optimize next token prediction, reducing redundant computations and speeding up the process (Touvron et al., 2023; Khadka, 2024).

All these state-of-the-art models have several variations and successive versions are released with improvements. Different (and bigger) training corpuses and an increase in their parameters, lead to better performance and improved capabilities. All the state-of-the-art models are summarized in Table 4, where we observe the quick evolution in the last years going from the 110M parameter of the early models in 2018, to the 405B parameters of the LLaMa 3.1 just released in July 2024. This increase in parameters also rely on bigger training corpuses, training time and small variations in transformer configurations that allow to improve the performance for specific tasks, such as text generation in GPT. Overall, the state-of-the-art is quickly evolving with substantive improvements in a very short time, allowing for its application in task such as emotion classification with a substantial increase in accuracy and contextualization as showcase in the following section. For each model, Table 4 shows the year it was released, which organization who developed it, the number of parameters the model has, its main characteristics, information about the training data used to train the model, and its related reference.

Table 4 Main pre-trained transformers and their characteristics, parameters and training corpus.

Model	Year of Release	Developed By	Parameters	Characteristics	Training Corpus Information	Reference
BERT	2018	Google AI	Base: 110M Large: 340M	Bidirectional; understands context from both left and right of a word	BERT is trained on a combination of BookCorpus, plus English Wikipedia, which totals 16GB of uncompressed text	(Devlin et al., 2019)
GPT	2018	OpenAI	117M	Unidirectional; predicts next word in sequence	Trained on diverse internet text	(Radford et al., 2018)
GPT-2	2019	OpenAI	1.5B	Unidirectional; generates coherent text	Trained on diverse internet text ~40 GB of text data	(Radford et al., 2019)
RoBERTa	2019	Facebook AI	Base: 110M Large: 340M	Optimized BERT; removed NSP, larger mini-batches, more data	Trained on dataset 10 times larger than BERT's, including Common Crawl	(Liu et al., 2019)
XLNet	2019	Google & CMU	Base: 110M Large: 340M	Combines autoregressive and autoencoding; permutation-based training	Trained on Wikipedia, BooksCorpus, Giga5, ClueWeb, and Common Crawl with subword pieces 32.89B in total	(Yang et al., 2020)
GPT-3	2020	OpenAI	175B	Unidirectional; generates highly coherent and contextual text	Trained on a mixture of datasets including filtered Common Crawl, WebText2, Books1, Books2, and Wikipedia with a total of 499 tokens	(Brown et al., 2020)
GPT-4	2023	OpenAI	Hundreds of billions (estimated)	Improved reasoning and understanding; handles complex prompts	Mix of filtered Common Crawl, WebText2, Books, Wikipedia, and other high-quality sources. Trained for 13 epochs, 4.5 trillion tokens.	(OpenAI et al., 2024)

LLaMA	2023	Meta (Facebook AI)	7B, 13B, 30B, 65B	Focuses on efficiency and accessibility; high performance with fewer parameters	Trained on 1.4 trillion tokens from English CommonCrawl, C4, Github, Wikipedia, Gutenberg and Books3, ArXiv and Stack Exchange	(Touvron et al., 2023)
LLaMa 3.1	2024	Meta (Facebook AI)	8B, 70B, and 405B	Enhanced training techniques, increased parameter sizes, more robust handling of diverse tasks from LLaMa	Over 15 trillion tokens (exact details of the training dataset mix released)	https://github.com/huggingface/blog/blob/main/llama31.md

Source: Authors' own elaboration

3.3.5. EXAMPLES OF DEEP LEARNING CLASSIFIERS APPLIED TO EMOTION DETECTION

We have conducted a comprehensive and thorough analysis of various models and techniques utilized for emotion detection. This detailed examination has provided valuable insights into the different methodologies employed to recognize and interpret emotional states from textual data. In this literature review, we explored a range of approaches, from traditional machine learning algorithms to advanced deep learning models, assessing their strengths, limitations, and applicability to different contexts based on their architecture and functioning principles. We will delve deeper into the practical applications of these models, specifically focusing on how they have been employed in real-world emotion detection scenarios. Examining how these models have been adapted and optimized for emotion detection tasks, considering factors such as feature extraction, data representation, and the integration of contextual information. We will analyse the performance of various methods, including individual techniques and ensemble approaches that combine multiple models to improve accuracy and robustness.

Xia & Zhang (2018) explore the three DL models for emotion detection in text that we have characterized in this report CNN, RNN and LSTM. The CNN model demonstrated the highest accuracy in emotion classification tasks, excelling due to its local sensing and parameter sharing capabilities, making it particularly effective for detecting a range of emotions such as joy, anger, sadness, and surprise. In contrast, RNNs can process sequential data but struggle with long-distance dependencies due to gradient extinction, which can hinder their performance in capturing emotions over longer text spans. LSTMs were introduced to address these limitations, effectively capturing long-range dependencies and improving the detection of complex emotional nuances. Overall, the experimental results indicated that all three models outperformed traditional machine learning methods like Support Vector Machines, with CNN being the most suitable for text emotion analysis, showcasing superior classification accuracy across various emotional categories. CNN models are some of the most widely used models in emotion detection has illustrated in the reviews identified, as their popularity raised in the last decade. Shrivastava et al. (2019) apply this CNN model to emotion detection in multimedia text data. It starts with pre-trained word embeddings to transform words into dense vector representations. Convolutional layers are then applied to detect local patterns and features related to emotional content. An attention mechanism is integrated to focus on significant parts of the text, enhancing contextual understanding. The processed data passes through fully connected layers to predict emotions. This model, evaluated on a manually annotated TV show transcript corpus, outperformed LSTM networks and random forest classifiers, demonstrating its effectiveness in capturing emotional nuances in text.

However, with the emergence in the last years of transformers, many researchers in the field have tested their capabilities in text classification, particularly fine-tuning some of the pretrained model like BERT and GPT for the task of emotion detection. Basile et al. (2019) fine-tuned the BERT model to classify emotions by using a two-sentence pair input format, where the first and third conversational turns are treated as separate sentences, ignoring the second turn. Each token is processed to create embeddings, and the model is trained on these embeddings to predict the emotion of the conversation. Additionally, lexical normalisation is applied to standardise the input text before feeding it into the BERT model for improved performance. The BERT model achieved an F1-score of 0.7263, which was the highest performance among the individual models used by Basile et al. (2019). This indicates that BERT was particularly effective in understanding and classifying the emotions in human-chatbot conversations. Nediko (2023) presents a study on emotion detection in a code-switching setting, specifically focusing on Roman Urdu and English SMS text messages. The author proposes a task for multi-class emotion classification, using generative pretrained transformers (GPT), particularly ChatGPT, due to its robust multilingual capabilities. They leveraged prompt engineering and few-shot learning methodologies to improve performance. The approach outperformed their baseline XGBClassifier and the organizers' BERT-based model, achieving a macro F1 score of 0.7038 and accuracy of 0.7313, ranking fourth in the competition.

Some of the more recent models developed such as the Emotion-LLaMA (Cheng et al., 2024) introduce a sophisticated approach to multimodal emotion recognition and reasoning by employing a multi-task learning strategy that combines emotional reasoning and recognition, also widening the scope to multimodal approaches. Emotion-LLaMA training involves two main phases: pre-training with coarse-grained visual and audio data to align multimodal features with word embeddings, followed by multimodal instruction tuning on fine-grained datasets such as MERR, MER2023, and DFEW. This process enhances the model's capability to accurately recognize and reason about emotions. The implementation uses various encoders, including EVA for global visual data, MAE and VideoMAE for local and temporal visual features, and HuBERT for audio, with training performed on GPUs for efficiency. Emotion-LLaMA outperforms state-of-the-art models across multiple metrics, achieving top scores in F1, Unweighted Average Recall (UAR), and Weighted Average Recall (WAR) on diverse datasets. Qualitative analyses further highlight its superior ability to integrate and interpret multimodal information, resulting in more accurate emotion recognition. Overall, Emotion-LLaMA represents a significant advancement in the field of multimodal emotion understanding, offering a robust solution for enhancing human-computer interactions and other applications requiring nuanced emotional comprehension.

CONCLUSIONS

This deliverable provided the state of the art for emotion and sentiment detection in text analysis. By conducting a scoping umbrella review and complementary desk research we identified the psychological models of emotions used in the scientific literature, with the model of Ekman (1992) being the most common one. Thereafter, we identified pre-processing techniques to prepare the text data for the models since text typically needs some cleaning and simplification to be more understandable. Since the models cannot understand plain text, the input text needs to be transformed into numerical representations. Several traditional methods for achieving this were outlined. Such methods are typically combined with a traditional ML classifier to identify emotions in text. The scoping umbrella review show that a vast variety of techniques exist for emotion and sentiment detection. Initially, lexicon approaches were used which are less data-driven since it requires human input to work properly by for example identifying relevant keywords or linguistic rules. Subsequently, more data-driven methods were developed. The ML methods can learn complex relationships in text while maintaining explainability meaning that it is possible to understand why the models make a certain prediction. However, the biggest breakthroughs for accurately completing NLP tasks such as emotion and sentiment classification has been achieved using DL techniques that are based on neural networks that can learn non-linear relationships between words and how their importance for predicting the next word in a sentence. The LSTM models achieved specific success due to its long memory meaning that it can use words early (or late) in a text to understand the context of any piece of text. However, the state-of-the-art models that have revolutionized the NLP field are the transformers. Introducing attention mechanisms coupled with encode-decoder architectures, they have shown superior performance on a wide range of tasks. The latest advancements in this area are the LLMs which are excellent at predicting the next word in a sentence by training on enormous amount of data and with trillions of parameters that are fine-tuned to capture the

Funded by the European Union

This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

underlying dimensions for understanding text. A great advantage of the transformers is that they are pre-trained, meaning that the models are available with the parameters already fine-tuned to understand text. Therefore, users of the models do not need to train the model from scratch but need only prompt it with relevant examples to fine-tune it to the specific context. In our case, this means training the transformers, that already understand text very well, for emotion classification.

The next step is to use the state-of-the-art in the methodology and subsequent social media analysis of emotions in WP3.

Funded by the European Union

This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

REFERENCES

Acheampong, F. A., Nunoo-Mensah, H., & Chen, W. (2021). Transformer models for text-based emotion detection: A review of BERT-based approaches. *Artificial Intelligence Review*, *54*(8), 5789-5829. https://doi.org/10.1007/s10462-021-09958-2

Agüero-Torales, M. M., Salas, J. I. A., & López-Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 107, 107373.

Alqahtani, G., & Alothaim, A. (2022). Predicting emotions in online social networks: Challenges and opportunities. *Multimedia Tools and Applications*, *81*(7), 9567-9605. https://doi.org/10.1007/s11042-022-12345-w

Alslaity, A., & Orji, R. (2024). Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. Behaviour & Information Technology, 43(1), 139-164.

Alswaidan, N., & Menai, M. E. B. (2020). A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, *62*(8), 2937-2987. https://doi.org/10.1007/s10115-020-01449-0

Basile A, Franco-Salvador M, Pawar N, Štajner S, Chinea Rios M, Benajiba Y (2019) Combined neural models for emotion classification in human chatbot conversations. In: Proceedings of the 13th international workshop on semantic evaluation, 2019. Association for Computational Linguistics, Minneapolis, pp 330–334

Belbasis, L., Bellou, V., & Ioannidis, J. P. (2022). Conducting umbrella reviews. BMJ medicine, 1(1).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Cambria, E., Livingstone, A., & Hussain, A. (2012). The hourglass of emotions. In Cognitive behavioural systems: COST 2102 international training school, dresden, Germany, February 21-26, 2011, revised selected papers (pp. 144-157). Springer Berlin Heidelberg.

Casola, S., Lauriola, I., & Lavelli, A. (2022). Pre-trained transformers: An empirical comparison. *Machine Learning with Applications*, *9*, 100334. https://doi.org/10.1016/j.mlwa.2022.100334

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. (2024). *Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction Tuning* (arXiv:2406.11161). arXiv. http://arxiv.org/abs/2406.11161

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Deng, J., & Ren, F. (2021). A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, *14*(1), 49-67.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Dictionary, M. W. (2002). Merriam-webster. *On-line at http://www. mw. com/home. html*, *8*(2), 23.

Drus, Z., & Khalid, H. (2019). Sentiment analysis in social media and its application: Systematic literature review. Procedia Computer Science, 161, 707-714.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

Funded by the European Union

This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

Eskandar, S. (2023, abril 26). *Exploring Feature Extraction Techniques for Natural Language Processing*. https://medium.com/@eskandar.sahel/exploring-feature-extraction-techniques-for-natural-language-processing-46052ee6514#:~:text=In%20natural%20language%20processing%20(NLP,its%20own%20strengths%20and%20weaknesses.

Garg, S., Panwar, D. S., Gupta, A., & Katarya, R. (2020, November). A literature review on sentiment analysis techniques involving social media platforms. In 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) (pp. 254-259). IEEE.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
Horev, R. (2018, noviembre 10). BERT Explained: State of the art language model for NLP. *Towards Data Science*. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

Khadka, P. (2024, abril 10). *LLaMA explained!* https://medium.com/@pranjalkhadka/llama-explained-a70e71e706e9

Kher, D., & Passi, K. (2022). Multi-label Emotion Classification using Machine Learning and Deep Learning Methods: *Proceedings of the 18th International Conference on Web Information Systems and Technologies*, 128-135. https://doi.org/10.5220/0011532400003318

Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I. (2023). A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. Artificial Intelligence Review, 56(12), 15129-15215.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. http://arxiv.org/abs/1907.11692

Lövheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. Medical hypotheses, 78(2), 341-348.
Munn, Z., Peters, M. D., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC medical research methodology, 18, 1-7.
Murthy, A. R., & Kumar, K. A. (2021, March). A review of different approaches for detecting emotion from text. In IOP Conference Series: Materials Science and Engineering (Vol. 1110, No. 1, p. 012009). IOP Publishing.
Nandwani, P., & Verma, R. (2021). A review on sentiment analysis and emotion detection from text. Social network analysis and mining, 11(1), 81.
Nedilko, A. (2023, July). Generative pretrained transformers for emotion detection in a code-switching setting. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis* (pp. 616-620).
OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Ortony A, Clore GL, Collins A (1988) The cognitive structure of emotions. Cambridge University
Parvin, T., & Hoque, M. M. (2021). An Ensemble Technique to Classify Multi-Class Textual Emotion. *Procedia Computer Science*, *193*, 72-81. https://doi.org/10.1016/j.procs.2021.10.008

Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., … & Yu, S. (2022). A survey on deep learning for textual emotion analysis in social networks. *Digital Communications and Networks*, *8*(5), 745-762.

Funded by the European Union

This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

Pereira, R., Mendes, C., Ribeiro, J., Ribeiro, R., Miragaia, R., Rodrigues, N., ... & Pereira, A. (2024). Systematic Review of Emotion Detection with Computer Vision and Deep Learning. *Sensors*, *24*(11), 3484.
Plutchik R (1980) Emotion. A psychoevolutionary. Synth
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.
Saffar, A. H., Mann, T. K., & Ofoghi, B. (2023). Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, *137*, 104258.
Sahoo, C., Wankhade, M., & Singh, B. K. (2023). Sentiment analysis using deep learning techniques: a comprehensive review. *International Journal of Multimedia Information Retrieval*, *12*(2), 41.
Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). Emotion detection in text: a review. *arXiv preprint arXiv:1806.00674*.
Sendari, S., Zaeni, I. A. E., Lestari, D. C., & Hariyadi, H. P. (2020). Opinion Analysis for Emotional Classification on Emoji Tweets using the Naïve Bayes Algorithm. *Knowledge Engineering and Data Science*, *3*(1), 50-59. https://doi.org/10.17977/um018v3i12020p50-59

Shaver, P., Schwartz, J., Kirson, D., & O'connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, *52*(6), 1061.
Shrivastava K, Kumar S, Jain DK (2019) An effective approach for emotion detection in multimedia text
data using sequence-based convolutional neural network. Multimed Tools Appl 78:29607–29639
Scherer KR (2005) What are emotions? And how can they be measured? Soc Sci Inf 44:695–729
Singh, N. K., Tomar, D. S., & Sangaiah, A. K. (2020). Sentiment analysis: A review and comparative analysis over social media. *Journal of Ambient Intelligence and Humanized Computing*, *11*(1), 97-117. https://doi.org/10.1007/s12652-018-0862-8

Stankovic, L., & Mandic, D. (2022). *Convolutional Neural Networks Demystified: A Matched Filtering Perspective Based Tutorial* (arXiv:2108.11663). arXiv. http://arxiv.org/abs/2108.11663

Susanto, Y., Livingstone, A. G., Ng, B. C., & Cambria, E. (2020). The hourglass model revisited. IEEE Intelligent Systems, 35(5), 96-102.
Thoyyibah, T., Haryono, W., Zailani, A. U., Djaksana, Y. M., Rosmawarni, N., & Arianti, N. D. (2023). Transformers in Machine Learning: Literature Review. Jurnal Penelitian Pendidikan IPA, 9(9), 604-610.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). arXiv. http://arxiv.org/abs/2302.13971

Udochukwu O, He Y (2015) A rule-based approach to implicit emotion detection in text. In: Biemann C, Handschuh S, Freitas A, Meziane F, Métais E (eds) Natural language processing and information systems. Lecture notes in computer science. Springer, Cham, pp 197–203
Ullah, A., Khan, S. N., & Nawi, N. M. (2023). Review on sentiment analysis for text classification techniques from 2010 to 2021. *Multimedia Tools and Applications*, *82*(6), 8137-8193.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. http://arxiv.org/abs/1706.03762

Funded by the European Union
This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

Vora, P., Khara, M., & Kelkar, K. (2017). Classification of tweets based on emotions using word embedding and random forest classifiers. *International Journal of Computer Applications*, *178*(3), 1-7.
Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., ... & Zhang, W. (2022). A systematic review on affective computing: Emotion models, databases, and recent advances. Information Fusion, 83, 19-52.
Xia, F., & Zhang, Z. (2018). Study of text emotion analysis based on deep learning. *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2716-2720. https://doi.org/10.1109/ICIEA.2018.8398170

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335-4385.
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (arXiv:1906.08237). arXiv. http://arxiv.org/abs/1906.08237
Zhao, Q. (2021). Social emotion classification of Japanese text information based on SVM and KNN. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-021-03034-x

Funded by the European Union

This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

ACRONYM	FULL NAME
D	Deliverable
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
DPO	Data Protection Officer
EC	European Commission
EASME	The Executive Agency for Small and Medium-sized Enterprises
GA	Grant Agreement
GDPR	General Data Protection Regulation
PC	Project Coordinator
WP	Work Package
TL	Task Leader
TF-IDF	Term Frequency-Inverse Document Frequency
DoA	Description of Action
PDMP	Personal Data Management Plan
SES	Socioeconomic status
SQM	Scientific and Quality Manager
POS	Part of Speech
PM	Person month
RNN	Recurrent Neural Network
LSTM	Long-short Term Memory
M	Month
ML	Machine learning
NLP	Natural-Language-Processing
NER	Named Entity Recognition

Funded by the European Union

This project has received funding from the European Union under the Horizon Europe Research & Innovation Programme (Grant Agreement no. 101132698 ENCODE).

ENCODE

Funded by
the European Union

This project has received funding from the European Union under
the Horizon Europe Research & Innovation Programme
(Grant Agreement no. 101132698 ENCODE).