ALGERIAN DEMOCRATIC AND POPULAR REPUBLIC

MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH

KASDI MERBAH UNIVERSITY OUARGLA

FACULTY OF NEW INFORMATION AND COMMUNICATION TECHNOLOGIES

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

THESIS SUBMITTED IN CANDIDACY FOR A MASTER DEGREE IN COMPUTER SCIENCE

OPTION FUNDAMENTAL COMPUTING

BY

BRIKI ZAHRAT EL HOUDA & HALIMI CHAHED ROUDAUNA

THEME

AN ENSEMBLE CLUSTERING-BASED APPROACH fOR EMOTION DETECTION IN YOUTUBE COMMENTS: A CASE STUDY ON THE GAZA WAR

SUPERVISOR: Mrs. SAADI WAFA

CO-SUPERVISOR: DR. MEZATI MESSAOUD

ACADEMIC YEAR: 2024/2025

Acknowledgments

First and foremost, we praise and thank Allah Almighty for His grace and mercy. We would like to extend our sincere gratitude and appreciation to **Mrs. Saadi Wafa** for her invaluable guidance, insightful advice, and continuous support throughout the preparation of this work. Her constant encouragement was a fundamental motivation behind the successful completion of this thesis.

We also express our deep gratitude to **Mr. Mezati Messaoud** for his kind advice and assistance, which greatly contributed to the advancement of this research.

Finally, we convey our heartfelt thanks and appreciation to our families **the Halimi and Briki families** for their unwavering support, patience, and encouragement throughout our academic journey and beyond. Their belief in us was a strong driving force behind this achievement.

Dedication

To those who taught me that words have impact, and that knowledge carries a message, To my dear father, who instilled in my heart strength and resilience, and was always a steadfast support. To my beloved mother, the source of tenderness, whose prayers were the light that illuminated my path in the darkest of moments.

To my brothers and sisters, companions of the journey and unwavering support, those who shared the dream and believed in me despite all challenges.

To my loyal friends, who were a soothing balm in times of fatigue, And to everyone who told me, "You can do it," those words left an unforgettable mark. To my brother's wife, whose presence brought warmth and strength to our lives. And to my study companion, HALIMI CHAHED ROUDAUNA, who shared with me moments of hardship, effort, and perseverance you were a true friend and a solid support throughout this academic journey. Thank you from the heart.

To every teacher who guided me, To every academic who opened the doors of research before me and planted in me a love for knowledge.

To the martyrs of Gaza, those whose bodies departed while their spirits continue to soar in the sky of dignity, To the wounded, the imprisoned, and the steadfast mothers, To a people who have not been defeated by pain, nor broken by siege. I dedicate this work not only out of loyalty, but in recognition that knowledge carries your cause, That a word can be a bullet, and that research can be a stand.

To everyone who carries Palestine in their heart. This dedication is from me to you with boundless gratitude.

Briki zahrat elhouda

The Messenger of Allah said: "He who does not thank people, does not thank Allah." True are the words of the Prophet .

All praise is due to Allah outwardly and inwardly, in the beginning and the end, for His countless blessings, and for His support and guidance in completing this humble work. To Him belongs all praise, as befits the majesty of His Face and the greatness of His Sovereignty.

To the soul of my beloved father, Halimi Miloud, may Allah have mercy on him. Though his body is gone, his spirit remains a companion to my prayers. His love still pulses in my heart, and his wise words continue to illuminate my path. I pray that this work serves as ongoing charity in the scale of his good deeds.

To my dear mother, the endless source of giving, and the unwavering support of my heart, no words I write can ever do you justice. It is through your prayers and your contentment that I was able to reach this milestone.

To my siblings: Bachir, Moussa, Ahmed, Souhaib, Rawnak, and Tadj Edinne, the heartbeat of our home and the most beautiful gifts Allah has granted me. You have always been a source of support and joy. My deepest thanks and love to each of you.

To my precious family, my generous uncle, my kind aunt, my beloved maternal aunt, my sisters-in-law, and everyone who supported me with a kind word, a prayer, or a sincere smile, you all have an unforgettable share in this achievement.

And to my study companion, Zahrat ElHouda Biriki, who shared with me the moments of exhaustion, effort, and striving, you were truly a wonderful friend and a steadfast support on this academic path. Thank you from the depths of my heart.

O Allah, make this work sincerely for Your noble face, beneficial to Your servants, and a source of goodness in this life and the Hereafter.

HALIMI CHAHED ROUDAUNA

Abstract

Emotion detection from text has become a critical research direction in natural language processing, particularly with the rise of emotionally charged and multilingual content on platforms like YouTube. This study focuses on unsupervised emotion detection from short YouTube comments related to the Gaza war, using data in both English and Arabic. It explores the role of emojis and different levels of textual "representation at the word and sentence levels" in enhancing emotional interpretation. By employing ensemble clustering techniques, including co-association matrices and the Multi-Metric Genetic Algorithm (MM-GA), the research aims to uncover latent emotional structures without relying on pre-labeled data.

The results demonstrate that sentence-level representations, especially when combined with emojis, significantly outperform word-level representations in terms of clustering quality, achieving higher scores on evaluation metrics such as Silhouette, Calinski-Harabasz, and Davies-Bouldin indices. Emojis emerge as strong emotional indicators, In particular, the Silhouette Score reached 0.3367 for Arabic comments, while for English comments, the score reached 0.3130, indicating strong cluster separation and cohesion. Emojis emerge as strong emotional indicators, particularly in Arabic comments, which often exhibit linguistic ambiguity. Comparative analysis shows that MM-GA and co-association-based clustering methods provide more stable and coherent performance than majority voting or traditional Mirkin-based approaches. Overall, the findings confirm the effectiveness of ensemble clustering in detecting emotional patterns from unstructured and diverse YouTube data. The proposed framework offers a scalable and language-aware solution for unsupervised emotion detection, particularly suited to multilingual and politically sensitive contexts.

Keywords: Emotion Detection, Ensemble Clustering, Natural Language Processing (NLP), YouTube Comments, Unsupervised Learning, Arabic Dialects, Sentence Representation, Emojis, MM-GA.

الملخص

أصبح اكتشاف المشاعر من النص اتجاهًا بحثيًا بالغ الأهمية في مجال معالجة اللغة الطبيعية، لا سيما مع تزايد المحتوى العاطفي ومتعدد اللغات على منصات مثل يوتيوب. يركّز هذا البحث على الكشف غير المراقب للمشاعر في تعليقات يوتيوب القصيرة المرتبطة بحرب غزة، باستخدام بيانات باللغتين الإنجليزية والعربية. وتستكشف الدراسة دور الإيموجي ومستويات التمثيل المختلفة للنصوص على مستوى الكلمات والجمل في تعزيز فهم المشاعر.

من خلال توظيف تقنيات التجميع الجماعي، بما في ذلك مصفوفة الترابط (co-association) وخوارزمية الجينات متعددة المعايير (MM-GA)، يسعى هذا العمل إلى الكشف عن البُنى العاطفية الكامنة دون الاعتماد على بيانات معنونة مسبقًا.

تشير النتائج إلى أن تمثيلات الجمل، لا سيما عند دمجها مع الإيموجي، تتفوّق بوضوح على تمثيلات الكلمات من حيث جودة التجميع، محققة نتائج أعلى على مؤشرات التقييم مثل Silhouette و Calinski-Harabasz و Davies-Bouldin. وقد بلغ مؤشر Silhouette في أفضل حالة 0.3367 للتعليقات العربية و 0.3130 للتعليقات الإنجليزية ، مما يعكس تكوين عناقيد عاطفية متماسكة وواضحة.

كما أثبتت الإيموجي فعاليتها كمؤشرات عاطفية، خاصة في التعليقات العربية التي غالبًا ما تتسم بالغموض اللغوي. وتُظهر المقارنات أن خوارزميات MM-GA وتقنيات التجميع القائمة على الترابط توفر أداءً أكثر استقرارًا وتماسكًا مقارنةً بطرق التصويت أو خوارزمية Mirkin التقليدية.

بوجه عام، تؤكّد النتائج فعالية التجميع الجماعي في اكتشاف أنماط المشاعر من البيانات النصية غير المنظمة والمتعددة اللغات. ويقدّم الإطار المقترح حلاً قابلاً للتوسعة وواعياً باللغات لاكتشاف المشاعر بشكل غير مراقب، وهو مناسب للسياقات متعددة اللغات والحساسة سياسيًا.

الكلمات المفتاحية: اكتشاف المشاعر، التجميع الجماعي، معالجة اللغة الطبيعية، تعليقات يوتيوب، التعلم غير المراقب، اللهجات العربية، تمثيل الجمل، الإيموجي، MM-GA.

Résumé

La détection des émotions à partir de texte constitue un axe de recherche essentiel en traitement automatique des langues, notamment avec la prolifération de contenus émotionnels et multilingues sur des plateformes telles que YouTube. Cette étude porte sur la détection non supervisée des émotions dans des commentaires courts publiés sur YouTube en lien avec la guerre de Gaza, en utilisant des données en anglais et en arabe. Elle examine le rôle des émojis ainsi que les différents niveaux de représentation textuelle — au niveau des mots et des phrases — dans l'amélioration de l'interprétation émotionnelle.

En s'appuyant sur des techniques de clustering ensembliste, incluant les matrices de co-association et l'algorithme génétique multi-métrique (MM-GA), la recherche vise à révéler les structures émotionnelles latentes sans recourir à des données annotées.

Les résultats montrent que les représentations au niveau des phrases, en particulier lorsqu'elles sont combinées avec des émojis, surpassent nettement celles au niveau des mots en termes de qualité de regroupement. L'évaluation a été réalisée à l'aide de plusieurs métriques telles que Silhouette, Calinski-Harabasz et Davies-Bouldin, cependant, le coefficient de Silhouette a été principalement retenu pour guider l'analyse. Les meilleures performances ont été obtenues avec un score de Silhouette de 0,3367 pour les commentaires en arabe et de 0,3130 pour les commentaires en anglais, ce qui indique une bonne séparation et cohésion des clusters émotionnels.

Les émojis apparaissent comme des indicateurs émotionnels puissants, notamment dans les commentaires en arabe, souvent marqués par une certaine ambiguïté linguistique. Les analyses comparatives révèlent que les approches basées sur le MM-GA et les matrices de co-association offrent des performances plus stables et cohérentes que les méthodes de vote majoritaire ou les approches classiques basées sur Mirkin.

En somme, les résultats confirment l'efficacité du clustering ensembliste dans la détection des émotions à partir de données YouTube non structurées et hétérogènes. Le cadre proposé constitue une solution extensible et sensible aux spécificités linguistiques pour la détection non supervisée des émotions, particulièrement adaptée aux contextes multilingues et politiquement sensibles.

Mots-clés : Détection des émotions, Clustering ensembliste, Traitement automatique des langues (TAL), Commentaires YouTube, Apprentissage non supervisé, Dialectes arabes, Représentation de phrases, Émojis, MM-GA.

Contents

List of Figures

List of Tables

General Introduction

In the digital age, social networks have become vital platforms where people express their opinions, thoughts, and emotions in real time. These platforms generate vast amounts of user-generated content daily, making them an invaluable source of data to understand collective sentiments and psychological trends. Among these platforms, YouTube has a unique position due to its visual content and the richness of its comment sections, which often reflect strong emotional and political engagement, especially during times of crisis and war.

One such context is the ongoing conflict in Gaza, which has spurred massive global interaction and emotional discourse online. The comments on related videos often contain raw, unfiltered emotional expressions that provide a valuable window into public sentiment. However, extracting and analyzing emotions from such short, noisy, and multilingual text data remains a significant challenge, especially in the absence of labeled datasets.

Emotion detection, also known as emotion recognition, involves identifying and classifying emotions such as joy, sadness, anger, fear, and others within textual content. Although supervised machine learning techniques require annotated datasets to train accurate models, acquiring such datasets, particularly for Arabic and politically sensitive content, is time consuming and often impractical. This limitation motivates the need for unsupervised approaches that can automatically discover patterns in unlabeled data.

In our study, we propose an unsupervised framework for emotion detection in short texts by leveraging ensemble clustering techniques. Unlike individual clustering algorithms, ensemble clustering combines the strengths of multiple algorithms, such as KMeans, DBSCAN, Agglomerative, and Spectral Clustering, and integrates their outputs using a co-association matrix or voting-based consensus, yielding more robust and accurate results.

To represent the semantic content of the comments, we use transformer-based language models, specifically AraBERT for Arabic comments and BERT for English. We also apply dimensionality reduction (e.g., PCA) to optimize clustering performance. The resulting clusters are labeled us-

ing the Ekman emotion model, which defines six basic emotions: happiness, sadness, fear, anger, surprise, and disgust.

A unique aspect of our work is the exploration of the role of emojis and keywords in enhancing emotion detection. Emojis, often overlooked in traditional models, carry significant emotional cues in short social media texts and can impact the detection process.

Through this research, we aim to contribute to the growing field of affective computing and social media analysis, offering new insights into unsupervised emotion detection from multilingual and politically sensitive content. Our findings are relevant for digital activism, mental health monitoring, public opinion analysis, and media content moderation.

This thesis is organized into four main chapters :

Chapter 1 presents the concept of emotion detection, starting with definitions, types, and models of emotions. Explores methods for detecting emotions, particularly in the context of social media platforms such as YouTube, and highlights the psychological impact of war, especially the war in Gaza, on online emotional expression.

Chapter 2 provides an overview of artificial intelligence and machine learning, with a particular focus on unsupervised learning techniques and clustering algorithms. It introduces the concept of ensemble clustering as a solution for improving clustering accuracy and robustness.

Chapter 3 explores the text-based emotion detection techniques. It examines emotion detection in short texts using various approaches, including keyword-based, rule-based, and machine learning techniques. It also discusses the use of clustering for emotion analysis and the associated challenges in processing short, informal social media texts.

Chapter 4 describes the conception and implementation of our proposed system. It details different steps of our system from data collection to ensemble clustering. This chapter also presents the evaluation metrics used and the results obtained from both Arabic and English datasets.

The thesis concludes with a general conclusion, summarizing our findings and presenting potential perspectives for future work.

CHAPTER 1

Emotion detection

1 Introduction

Every day, millions of individuals utilize social media; these sites allow users to publish, remark, and like big events, thereby expressing their emotions, ideas, and thoughts. In this chapter, we explore the context of emotion detection on social media. Beginning by defining key terms, we examine the different key elements and types of emotions, comparing them to feelings and moods. Different prominent models for classifying emotional expressions are presented. Methods for emotion detection are discussed, focusing on the techniques used to detect emotions from text. The specific context of emotion detection on platforms such as YouTube is also explored, highlighting the psychological effects of war. This foundation provides a deeper understanding of how emotions are detected, interpreted, and used in the digital world.

2 Definition of Emotion

Emotion [1], a term whose precise meaning psychologists and philosophers have contested over for more than a century. In its most literal sense, the Oxford English Dictionary defines emotion as "any agitation or disturbance of mind, feeling, passion; any vehement or excited mental state." We take emotion to refer to a feeling and its distinctive thoughts, psychological and biological states, and the range of propensities to act. [2].*'Emotions are a process, a particular kind of automatic evaluation influenced by our evolutionary and personal past, in which we sense that something important to our welfare is happening and a set of psychological changes and emotional behaviors begins to deal with the situation.'*[3]

3 Key Elements of Emotion

To better understand what emotions are, let's focus on their three basic elements, also known as the emotion process.

3.1 Subjective Experience

The subjective experience of emotion [1], how each emotion feels, is for some at the center of what an emotion is. This presumably includes physical sensations, and other feelings which are the consequence of feedback from the various response changes which occur uniquely for each emotion. Regrettably, most of what we know about subjective experience comes from questionnaires, filled out by people who are not having an emotion, trying to remember what it feels like. It is no easy matter to assess subjective experience, especially if what is wanted is something more than simply the amount of positive or negative emotion

3.2 Physiological Responses

Emotions prepare the body for action by simultaneously activating certain systems and deactivating others in order to prevent the chaos of competing systems operating at the same time, allowing for coordinated responses to environmental stimuli [4]. For instance, when we are afraid, our bodies shut down temporarily unneeded digestive processes, resulting in saliva reduction (a dry mouth); blood flows disproportionately to the lower half of the body; the visual field expands; and air is breathed in, all preparing the body to flee. However, one common misunderstanding many people have when thinking about emotions is the belief that emotions must always directly produce action [5] .This is not true. Emotion certainly prepares the body for action, but whether people actually engage in action is dependent on many factors, such as the context within which the emotion has occurred, the perceived consequences of one’s actions, and previous experiences.

3.3 Behavioral Responses

Emotions shape our conduct and serve as powerful motivators for future action. Many of us seek to feel satisfied, joyful, proud, or triumphant about our achievements. At the same time, we try to prevent powerful negative emotions. For example, once we have experienced disgust from drinking spoiled milk, we generally work very hard to avoid having those feelings again (e.g., checking the expiration date on the label before purchasing the milk, smelling the milk before drinking it, and

checking to see if the milk curdles in one's coffee before drinking it). As a result, emotions not only impact current acts but also serve as a powerful motivator for future behaviors. [6]

Emotions are communicated both vocally and non verbally, through facial expressions, voices, gestures, bodily postures, and movements. According to research [7], humans are very sensitive to emotional information sent by body language, even if we are not consciously aware of it. Humans continuously display their emotions when engaging with others. Emotions and expressions provide information to others about our feelings and intentions.

4 Types of Emotion

In emotional psychology, emotions are classified into different types based on their nature and impact. Emotions are divided into basic emotions, also known as primary emotions, and complex emotions, also known as secondary emotions, with each category having specific characteristics that contribute to our emotional responses.

4.1 Basic Emotions

Paul Ekman research [1] showed that people, regardless of their cultural background, express and recognize specific emotions through facial expressions in a similar way. his originally identified six basic emotions that are universally recognized: Happiness, Sadness, Fear, Anger, Surprise, Disgust.

4.2 Complex Emotions

Complex emotions have differing appearances and may not be as easily recognizable, such as grief, jealousy or regret. Complex emotions [8] are defined as *"any emotion that is an aggregate of two or more others."*Hate is considered a blend of fear, anger, and disgust. Basic emotions, however, are pure and innate. Other complex emotions include love, embarrassment, envy, gratitude, guilt, pride, worry, and many more. Complex emotions vary greatly in how they appear on a person's face and don't have easily recognizable expressions. Grief looks quite different between cultures and individuals. Some complex emotions, such as jealousy, may have no accompanying facial expression at all.

5 Comparison between Emotions and Feelings and Moods

Emotions are often confused with feelings and moods, but the three terms are not interchangeable. According to the American Psychological Association (APA), emotion [9] is defined as *'a complex reaction pattern, involving experiential, behavioral, and physiological elements.' Emotions are how individuals deal with matters or situations they find personally significant. Emotional experiences have three components: a subjective experience, a physiological response, and a behavioral or expressive response.*

Feelings arise from an emotional experience. Because a person is conscious of the experience, this is classified in the same category as hunger or pain. A feeling is the result of an emotion and may be influenced by memories, beliefs and other factors.

A mood [10] is described by the APA as *"any short-lived emotional state, usually of low intensity."* Moods differ from emotions because they lack stimuli and have no clear starting point. For example, insults can trigger the emotion of anger while an angry mood may arise without apparent cause.

6 Emotion models

Human emotions play a pivotal role in interpreting behavior, rendering them a fundamental subject of inquiry in media studies and content analysis. Psychological research has demonstrated that individuals react to specific stimuli through a set of basic emotions, such as joy, anger, sadness, fear, and disgust. These emotional responses can be systematically measured and analyzed using scientific methodologies, allowing researchers to evaluate affective reactions within media messages and communication contexts.

To facilitate a deeper understanding of emotional processes, scholars have proposed two principal models:

The Categorical Model: This model classifies emotions into distinct, discrete categories.

The Dimensional Model: This approach conceptualizes emotions along continuous dimensions, such as intensity, valence (positive or negative), and arousal level.

6.1 Categorical model

Placing emotions into discrete classes or categories is a component of the categorical model of emotions. Among the most notable are:

6.1.1 The Paul Ekman model

According to researcher Paul Ekman, there is a set of basic human emotions which include fear, surprise, anger, disgust, sadness, and happiness. Ekman argues that complex emotions emerge from combinations of these basic emotions, such as greed, lust, humiliation, regret, pride, and others. For example, fear combined with sadness may result in the feeling of regret.

6.1.2 The Robert Plutchik model

Plutchik's model is based on the idea of basic emotions that form the foundation of the human emotional experience At the core of this model lie eight primary emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. Plutchik represents these emotions in a circular format, where Each emotion is paired with its opposite emotion: joy versus sadness, anger versus fear, trust versus disgust, and surprise versus anticipation. The model adds another dimension of understanding by illustrating the intensity of each emotion

6.1.3 Orthony, Clore, and Collins (OCC) model

The OCC model is a structured scientific framework [11] aimed at understanding and classifying human emotions. This model analyzes and organizes the fundamental components that cause the emergence of various emotions. The model relies on classifying emotions based on three key elements: the source of emotion, its nature, and its impact on human behavior. Through this analysis, 22 different types of emotions have been identified, divided into six main categories.

6.2 Dimensional Emotion Models

A dimensional model is a sophisticated scientific framework for studying and quantifying human emotions. These theories move away from standard categorical classifications of emotions and toward a more complex, multidimensional approach.

6.2.1 Russell model

The model provides a structured scientific conceptualization of human emotions, arranging them in a circular, sequential pattern [12], Based on the premise that emotions are not entirely discrete but rather flow in a natural and logical sequence, this model highlights the gradual and interconnected evolution of emotions. Its effectiveness has been substantiated by numerous scientific studies, making it a valuable tool for understanding and interpreting the human emotional experience.

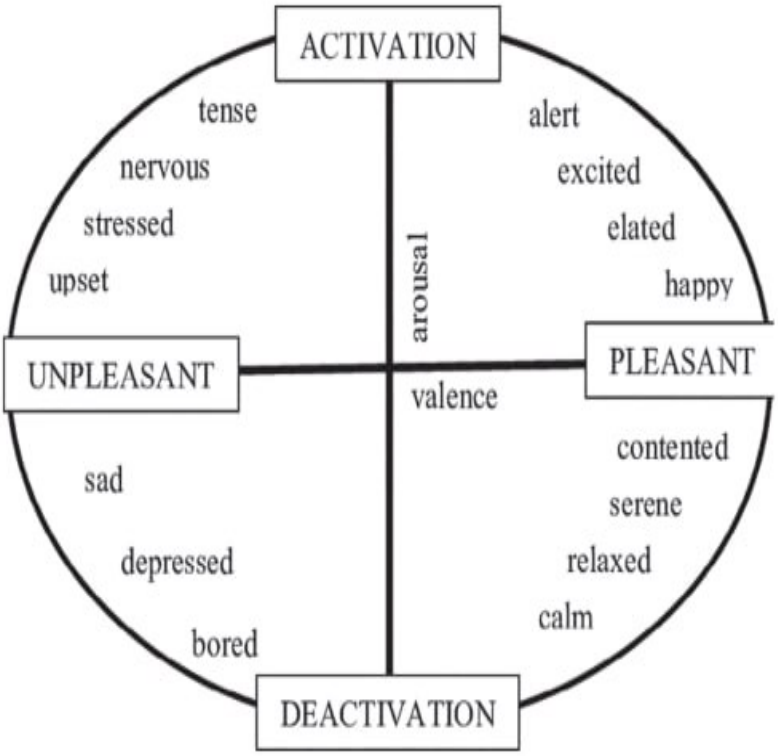


Figure 1.1: Russell model. [12]

6.2.2 Plutchik model

The Plutchik model [13] depicts emotions as a wheel of concentric circles, with primary emotions on the outside and the eight essential emotions on the inside. Deep feelings are derived from the eight fundamental emotions. The wheel depicts the level of relatedness between emotions based on their position on the wheel. Emotions are conveyed in opposite pairs: surprise versus anticipation, joy versus sadness, rage versus fear, and trust versus contempt.

Figure 1.2: Plutchik model.[13]

6.2.3 Russell and Mehrabian model

Mehrabian and Russell [14] proposed pleasure, arousal, and dominance as three distinct emotional dimensions for describing people's feelings. They defined pleasure as a continuum that ranged from extreme pain to extreme bliss, using terms such as happy unhappy, pleased-annoyed, and satisfied unsatisfied. Arousal was a mental activity that ranged from sleep to disagreeable pleasantness, while dominance was linked to feelings of control and restricted conduct. The term "relaxation" served as an indicator for all three aspects.

7 Emotion detection

Emotion detection [15][16] is considered one of the advanced branches of sentiment analysis in the field of informatics. It focuses on identifying and analyzing emotional states within written texts. This process holds particular importance in digital contexts, where emotions play a central role in shaping user interaction, especially on social media platforms [17]. Emotion detection systems rely on intelligent algorithms and Natural Language Processing (NLP) techniques to automatically

recognize, understand, and handle human emotions.

7.1 Methods of Emotion Detection

Emotion detection encompasses a diverse set of techniques that have evolved significantly with technological advancements, enhancing the ability of systems to recognize and interpret human emotional states. These methods include:

7.1.1 Facial recognition

Facial expression analysis [18] is a crucial scientific discipline that integrates the observation of facial features with the interpretation of emotional expressions. The human face serves as the primary channel through which emotions are externally conveyed. By analyzing facial muscle movements and expressions, emotional states can be identified. This field employs several methodologies, including facial expression encoding systems, and leverages computer vision and machine learning algorithms to detect and interpret subtle facial changes.

7.1.2 Speech analysis

Emotional cues can be inferred from audio signals by examining variations in pitch, intensity, and rhythm [18]. This method typically involves two core stages: extracting relevant acoustic features, and classifying them to identify associated emotional states. A variety of models, including statistical and deep learning approaches, are utilized to enhance the accuracy of speech based emotion detection.

7.1.3 Physiological sensors

Physiological signals [18], such as heart rate, skin conductance, and body temperature, reflect the body's biochemical responses to external stimuli. One of the most widely used signals in this context is the electrocardiogram (ECG), which allows for emotion detection by analyzing heart activity. Due to the complexity of such signals, advanced signal processing techniques and machine learning models are essential for accurate interpretation.

7.1.4 Textual Emotion Recognition

Text-based emotion recognition [18] is a rapidly evolving area within natural language processing. It involves analyzing written content such as social media posts, reviews, and messages to

determine the emotional state of the author. Techniques include lexicon based methods and knowledge enhanced neural networks that utilize semantic ontologies to improve detection precision.

7.2 Social Media and Emotion Detection

Social media plays a significant role in our modern lives, serving as an effective means of communication between individuals and a platform for expressing thoughts and emotions on a large scale. The widespread availability of smart devices enables users to seamlessly navigate between applications such as Facebook [19], Twitter [20] [21], YouTube [22], and Instagram. These platforms host vast amounts of real time, user generated textual data, reflecting opinions, interests, and emotional states. This makes social media a rich source for applying emotion detection techniques [17] [16] [23] to better understand collective moods, public attitudes, and emerging trends [24].

Emotion detection has emerged as a prominent field that leverages this digital momentum." *Given the central role social media plays in contemporary communication, researchers are now able to study a wide range of digital interactions between individuals, including public opinions on social and political issues. To accomplish this, it is essential to employ various techniques for detecting emotions and identifying affective states throughout these processes*"[25].

7.2.1 YouTube Platform

YouTube [22] provides a valuable medium for identifying emotional states through the analysis of user comments. Each day, millions of comments are submitted via the platform's Application Programming Interface (YouTube API v3), offering a vast and dynamic dataset for analysis. In response to this, numerous automated sentiment analysis [26] techniques have been developed by researchers to classify and interpret emotional expressions such as joy, sadness, anger, fear, and others. The accurate detection of these emotions on YouTube demonstrates significant potential across various applications, particularly with the continued advancement of computational linguistics and natural language processing technologies.

8 Psychological impacts of the War

Modern armed conflicts have, in many instances, seen an increase in the civilian toll compared to previous traditional wars. Traumatic experiences in armed conflicts include physical injury, close to death, imprisonment, inaccessibility to food, water, and healthcare, lack of shelter, forced displacement, separation from family, and others. Depression, anxiety, and posttraumatic stress disorder (PTSD) represent the main mental health problems identified in individuals affected by

armed conflict. Other impacts involve the social functioning and personality of affected individuals and their risk of substance abuse. The WHO [27] estimated that, in the situations of armed conflicts throughout the world, *"10% of the people who experience traumatic events will have serious mental health problems and another 10% will develop behavior that will hinder their ability to function effectively. The most common conditions are depression, anxiety and psychosomatic problems such as insomnia, or back and stomach aches"*

8.1 Recent Technique of Detecting Impacts Of the War

Techniques used to detect the psychological impact of war have evolved significantly with advances in artificial intelligence, data science, and behavioral psychology. Traditional methods, such as clinical interviews and questionnaires, are no longer sufficient. They have been augmented by advanced technologies that enable a deeper and faster analysis of psychological changes in individuals affected by war. Modern methods rely on artificial intelligence and machine learning techniques to analyze massive amounts of data derived from social media.

8.1.1 Social Media and Psychological Impact of War

- **Sentiment Analysis in Posts**

  Machine learning models are being used to analyze tweets, Facebook posts, and YouTube [**reddy_emotion_2023**] to detect emotions such as fear, anger, and sadness. Example: An analytical study of tweets about the Russian-Ukrainian war showed an increase in the expression of anxiety and depression among users.

- **Studying Linguistic Patterns in Digital Texts**

  Text analysis techniques are used to detect changes in writing style, such as frequent use of emotionally charged or negative expressions that reflect psychological distress. In the digital age, this also includes the use of abbreviations (for example, 'IDK', 'OMG'), acronyms, emojis, and autocorrected spelling ,all of which have become part of the patterns of digital language of users. These elements can signal anxiety, urgency, or emotional detachment and are analyzed to understand the psychological state of individuals posting during wartime.

- **Monitoring Digital Interactions and Behavior**

  Researchers analyze users engagement with war related content, such as sharing traumatic news or emotionally intense comments to assess psychological responses. For instance, a study published in Computers in Human Behavior found that specific social media activities

mediate the relationship between overall social media usage and psychological distress, with factors like media skepticism and interpersonal injustice acting as moderators (Linking components of social media usage to psychological distress: Integrating the situational theory of problem solving and social capital theory).

- **Emoji and hashtag analysis**

  Data science tools analyze popular emojis and hashtags (e.g., #GazaUnderAttack, #PrayForUkraine) during conflicts to understand emotional trends in online communities.

- **Video and Audio Analysis**

  AI powered tools analyze voice tone and facial expressions in videos to detect signs of distress, anxiety, or depression caused by war trauma. For instance, refugee testimonies on TikTok have been studied using machine learning models to detect emotional shifts through tone, pauses, and facial micro expressions. Moreover, TikTok's unique digital features such as short form videos, filters, voice overlays, and the "duet" function encourage expressive storytelling that mixes humor, fear, irony, and resilience.

8.2 Emotion Detection During The War

8.3 The War of Gaza

The Gaza Strip has been subjected to a devastating military assault since October 2023, resulting in severe consequences for the civilian population. This ongoing conflict poses a significant threat to the mental health of affected communities, particularly in the context of repeated trauma and forced displacement. Psychiatric morbidity is known to escalate under such conditions.

A study conducted in 2024 investigated the prevalence and risk factors associated with depression, anxiety, stress, and post traumatic stress disorder (PTSD) among young adult students from the Gaza Strip during the war [29]. The study involved 339 medical students, the majority of whom had experienced multiple displacements and reported the loss of relatives, colleagues, or friends. A significant portion also reported the loss of their homes and sources of income.

The findings revealed that 97.05% of participants exhibited at least mild depressive symptoms. Additionally, 84.37% and 90.56% reported symptoms of mild anxiety and stress, respectively. High levels of life dissatisfaction were also prevalent, with 63.40% meeting the criteria for PTSD. The incidence of psychiatric symptoms was significantly higher than baseline rates, and all participants diagnosed with PTSD presented with at least one comorbid psychiatric condition.

Further analysis indicated that living in a shelter and experiencing moderate or severe stress symptoms were significantly associated with depression. Predictors of moderate or higher anxiety included being female, losing a friend, experiencing moderate stress, and having PTSD. Likewise, moderate or higher levels of depression, anxiety, and the presence of PTSD predicted the occurrence of moderate or higher stress symptoms. Finally, both anxiety and stress at moderate or higher levels were strong predictors of PTSD.

8.3.1 Social media effect of war

A lot of researchers have conducted studies regarding the sentiment analysis during of conflit using various tetechniques and procedures,hich will be considered as part of the theoretical framework of this research, enabling afterward the construction of the conceptual framework of the study.For example, in September 2024, Hofmann et al. conducted a study [30] to explore the prevalence of hate speech (HS) and sentiment in YouTube video comments related to the Israel Palestine conflict by analyzing content from both public and private news sources. The study involved annotating 4,983 comments for HS and sentiment categories (neutral, pro Israel, and pro Palestine). Machine learning models were developed, demonstrating strong predictive capabilities with area under the receiver operating characteristic (AUROC) scores ranging from 0.83 to 0.90. These models were applied to YouTube comment sections extracted from both public and private sources, revealing a higher incidence of hate speech in public sources (40.4%) compared to private sources (31.6%).

Sentiment analysis revealed that the majority of comments were neutral in both source types, with more pronounced sentiments towards Israel and Palestine observed in public sources. This study highlights the dynamic nature of online discourse surrounding the Israel Palestine conflict and emphasizes the importance of moderating content in politically charged environments.

The June 2024 study by Ashagrew Liyih et al [31]. addresses the sentiment analysis of public opinion regarding the Hamas Israel war by analyzing a substantial dataset of 24,360 YouTube comments from prominent news sources such as BBC, WION, and Al Jazeera. The comments were classified by linguistic experts into three categories: positive, negative, and neutral. The researchers utilized natural language processing (NLP) methods and feature extraction techniques, including Word2Vec, FastText, and GloVe. They then experimented with several deep learning architectures, such as LSTM, BiLSTM, GRU, and a hybrid CNN BiLSTM model.

The hybrid CNN LSTM model, combined with Word2Vec embeddings, yielded the best performance, achieving an impressive accuracy of 95.73%. This study underscores the power of deep learning models in sentiment classification, especially when applied to politically sensitive topics like the Hamas Israel war. The effectiveness of Word2Vec embeddings in enhancing the model's

performance is notable, as it contributed to the high accuracy in identifying sentiments from large scale YouTube comment datasets. This research highlights the potential of advanced NLP techniques in understanding and analyzing public opinion, particularly in politically charged contexts.

In their study [32], Temel Eğinli and Özmelek Taş (2023) analyzed 10,000 tweets posted during the Russia Ukraine war using sentiment analysis and content analysis. Their results showed that 39.5% of the tweets expressed positive emotions, 45% were neutral, and only 15.5% conveyed negative sentiment. The study highlighted the tendency of users to express supportive and empathetic feelings while avoiding overly negative or violent language. This research underlines how social media platforms reflect public emotions during international conflicts and contribute to shaping collective discourse. In another [33]large-scale sentiment analysis study analyzed 603,552 English tweets and 1,664 Russian tweets related to the Russia Ukraine war, posted during the early phase of the conflict (January to March 2022). The researchers used DistilRoBERTa and XLM RoBERTa to classify emotions. English tweets were categorized into seven emotional states, with fear (32.08%) and anger (15.18%) being the most frequent. Russian tweets showed overwhelmingly negative polarity (86.83%). The study highlighted the prevalence of expressions supporting Ukraine, denouncing the war, and expressing concern over weapons and casualties. These findings confirm that social media users often turn to platforms like Twitter to process and express strong emotional responses during times of war. The authors suggest that future research could expand the dataset and segment tweets by country to better understand geographic sentiment differences.

8.3.2 Emotion Detection Research Handling The Effect Of The Gaza War To The World

Studies focusing on the impact of the Gaza conflict on social media have shown that these platforms play a crucial role in expressing emotions and political engagement during times of conflict. On 20 January 2025, Nabhani et al. conducted a study titled "Integrating Argumentation Features for Enhanced Propaganda Detection in Arabic Narratives on the Israeli War on Gaza." This study [34]represents a significant advancement in automated propaganda detection within Arabic texts, particularly in conflict driven contexts such as the Israeli Palestinian war. It integrates argumentation features including claims, premises, and major claims into machine learning models to improve the identification of propaganda techniques in Arabic media. The authors utilize finely annotated datasets and combine cross lingual and multilingual NLP approaches with GPT-4 based annotations, resulting in consistent performance improvements.

In June 2024, Younes conducted a study [35] to compare various Natural Language Processing (NLP) techniques for sentiment analysis on social media data related to the Gaza conflict. The study involved analyzing public comments to assess sentiments regarding the war. Different classi-

fication methods, including traditional machine learning, deep learning, and transfer learning, were employed. The results indicated that the majority of comments expressed negative sentiments towards the war. Notably, the DistilBERT classifier achieved the highest classification accuracy at 89%, slightly outperforming the LSTM model, which achieved an accuracy of 88%. The findings of this study contribute to advancing the field of sentiment analysis, providing valuable insights for future research on public sentiment in the context of global conflicts.

In late 2023, Lynette Ng, Adrian Lim, and Roy Lee developed a dataset titled "Love-Hate Dataset" [36] which is a multimedia and social media platform dataset comprising English language posts from Facebook and Instagram. These posts express sentiments related to "love" and "hate" during the Israel Hamas conflict between October 7 and December 31, 2023.The study revealed that the number of posts related to the war declined over time, indicating a decrease in interest. It also showed that "love" posts often included religious references and used emojis symbolizing hearts, peace, and listening, while "hate" posts included references to hostility and employed emojis representing sadness, warning, and surveillance. In an additional step, the researchers used the GPT-4 model to generate new Instagram posts based on this data. They found that the model predominantly produced general love messages accompanied by artistic images, without delving into the specifics of the conflict.This study aims to assist researchers in studying multimedia data and emotions on social media platforms during wars and represents an important contribution to the field of digital sentiment analysis in political conflict contexts.

9 Conclusion

Emotion recognition and research is an important and advanced field, particularly concerning social media and contemporary conflicts. Using different models to understand emotions, including categorical and dimensional models, provides a deeper understanding of the complexities associated with human emotional reactions. Recent advancements, such as machine learning and natural language processing, have significantly enhanced the accuracy of emotion recognition in short texts, including social media posts and YouTube comments. These technologies serve as a powerful resource for assessing the psychological impacts of major events such as wars, as social media platforms reflect the emotional states of those affected. In the next section, we briefly illustrate the existing Artificial Intelligence technique.

CHAPTER 2

Artificial Intelligence

1 Introduction

In recent years, Artificial Intelligence (AI) has emerged as a transformative force across various domains, from healthcare and finance to transportation and social media. The growing ability of machines to simulate aspects of human intelligence such as learning, reasoning, and decision making has opened new avenues for solving complex problems that were previously considered exclusive to human cognition. This chapter provides a comprehensive overview of artificial intelligence, with a particular emphasis on machine learning (ML), which constitutes the backbone of most modern AI systems.

The chapter begins by defining AI and its core objectives, followed by an exploration of machine learning and its primary types, including supervised, unsupervised, semi supervised, and reinforcement learning. Special attention is given to unsupervised learning techniques, particularly clustering methods, which play a pivotal role in discovering hidden patterns and structures in unlabeled data. The chapter concludes by introducing ensemble clustering, a powerful approach that leverages the strengths of multiple clustering algorithms to enhance the robustness and accuracy of clustering outcomes.

2 AI Definition

Kurzweil defined artificial intelligence as the art of creating machines that perform functions that require intelligence when performed by people. [37]

Most AI examples that you hear about today from chess playing computers to self driving cars,

rely heavily on deep learning and natural language processing. Using these technologies, computers can be trained to accomplish specific tasks by processing large amounts of data and recognizing patterns in the data.

3 Machine Learning

3.1 Definition

Machine learning (ML) is a branch of artificial intelligence (AI) focused on enabling computers and machines to imitate the way that humans learn, to perform tasks autonomously, and to improve their performance and accuracy through experience and exposure to more data.

4 Types of Machine Learning

Machine Learning [37] algorithms are mainly divided into four categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.

4.1 Supervised Learning

Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input output pairs [38]. It uses labeled training data and a collection of training examples to infer a function. Supervised learning [39] is carried out when certain goals are identified to be accomplished from a certain set of inputs, i.e., a task-driven approach.

4.1.1 Classification

The classification [40] is defined as the process of assigning one or some among the predefined categories to each item. The binary classification, where each item is classified into one of two categories, is mentioned as the simplest classification type. The binary classification is expanded into a multiple classification by predefining more categories. If the given task is the soft classification, in which more than one category is allowed to be assigned to each item, the multiple classification is decomposed into binary classifications. This section is intended to describe briefly the binary classification, the multiple classification, and its decomposition into binary classifications.

4.1.2 Regression

Regression [40] is a type of machine learning technique used to predict a continuous value based on input data. The goal of regression is to find the relationship between the target variable (the one being predicted) and the independent variables (the inputs). It is used in a wide range of applications, such as predicting house prices, estimating product demand, or forecasting weather conditions.

4.2 Semi-supervised learning

Semi-supervised learning [38] can be defined as a hybridization of the supervised and unsupervised methods, as it operates on both labeled and unlabeled data. Thus, it falls between learning “without supervision” and learning “with supervision." In the real world, labeled data could be rare in several contexts, and unlabeled data are numerous, where semi supervised learning is useful.[39] The ultimate goal of a semi supervised learning model is to provide a better outcome for prediction than that produced using the labeled data alone from the model.

4.3 Reinforcement learning

Reinforcement learning [41] is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficien, i.e., an environment-driven approach. This type of learning is based on reward or penalty, and its ultimate goal is to use insights obtained from environmental activists to take action to increase the reward or minimize the risk. It is a powerful tool for training AI models that can help increase automation or optimize the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing and supply chain logistics; however, it is not preferable to use it for solving the basic or straightforward problems.

4.4 Unsupervised learning

Unsupervised learning analyzes unlabeled datasets without the need for interference, i.e., a data-driven process. This is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes.

4.4.1 Dimension reduction

Dimensionality reduction [42]is a technique used to reduce the number of variables or features in a dataset while preserving as much relevant information as possible. High dimensional data often poses challenges such as increased computational cost, overfitting, and the curse of dimensionality. Popular dimensionality reduction methods include Principal Component Analysis (PCA) and t distributed Stochastic Neighbor Embedding (t-SNE). These techniques transform the data into a lower dimensional space, making it easier to visualize and process. In emotion detection tasks, dimension reduction helps in simplifying complex textual embeddings while retaining the most informative features for clustering or visualization.

4.4.2 Clustering

Clustering [43] is a popular unsupervised learning technique that is designed to group objects or observations together based on their similarities. Clustering has a lot of useful applications, such as market segmentation, recommendation systems, exploratory analysis, and more.

For instance, clustering algorithms can analyze text data to identify and group that express similar emotions in social media, aiding in understanding public sentiment or trends.

Clustering algorithms group data points into clusters based on their similarities or differences. The goal is to identify natural groupings in the data. Clustering algorithms are divided into multiple types based on the methods they use to group data.

4.4.2.1 Types of Clustering Methods

1. **Centroid-Based Methods:**

   Centroid-based clustering algorithms [44] represent data points using central vectors, which do not necessarily belong to the original dataset. The fundamental principle behind these algorithms is the computation of distance measures between data instances, typically using metrics such as Euclidean, Manhattan, or Minkowski distances. Generally, the mean is used with the Euclidean distance to determine the centroid of a cluster, while the median is commonly applied with the Manhattan distance. Additionally, the steepest descent method can be employed in optimization contexts to improve convergence when minimizing objective functions in certain clustering or machine learning algorithms.

2. **Distribution-based Methods:**

   Distribution-based clustering methods [45] assume that data is generated from a mixture of underlying probability distributions, such as Gaussian distributions. One of the most well known examples is the Gaussian Mixture Model (GMM), which uses the Expectation Maximization (EM) algorithm to estimate the parameters of the distributions. Each cluster is associated with a probability distribution, and each data point is assigned a probability of belonging to each cluster.

   These methods are powerful when the data conforms to known statistical properties but can be sensitive to initialization and prone to overfitting if not regularized properly.

3. **Connectivity based methods:**

   Connectivity-based (or hierarchical) [46] clustering relies on the principle that objects that are closer to each other are more related than those further apart. These methods build clusters step by step either in a bottom-up approach (agglomerative) or a top-down approach (divisive). The result is usually represented in a dendrogram.

   Agglomerative hierarchical clustering is widely used due to its simplicity and interpretability. It does not require the number of clusters to be defined beforehand, which is a major advantage in exploratory data analysis.

4. **Density Based Methods:**

   Density based clustering [47] identifies clusters as dense regions in the data space separated by regions of lower density. The most prominent algorithm in this category is DBSCAN (Density Based Spatial Clustering of Applications with Noise). It groups together points that are closely packed and marks points in low density regions as outliers.

   This method is particularly effective in identifying clusters of arbitrary shape and is robust to noise, making it suitable for real world applications such as spatial data analysis or social media sentiment clustering.

4.4.2.2 Comparison between algorithms of clustering

Clustering Algorithm	Advantages	Disadvantages
K-means	• Simple and fast to implement. • Efficient for large datasets. • Produces reproducible results with fixed initialization. • Performs well with linearly separable (spherical) clusters.	• Requires specifying the number of clusters in advance. • Cannot handle non linear or irregularly shaped clusters. • Sensitive to initial centroid selection.
Agglomerative Clustering	• No need to predefine the number of clusters. • Allows visualization through dendrograms. • May capture non linear relationships in some cases.	• Computationally expensive for large datasets. • Merges/splits are irreversible. • Less effective for highly non linear structures.
Spectral Clustering	• Effective for complex, non convex, and non linear structures. • Utilizes similarity graphs for flexible modeling. • Good performance on non linearly separable data.	• Requires costly eigen decomposition. • Needs predefined number of clusters. • Not scalable to very large datasets.
DBSCAN	• Detects arbitrarily shaped (non linear) clusters. • Does not require predefined number of clusters. • Robust to noise and outliers. • Suitable for density based data.	• Sensitive to parameter settings (ε, MinPts). • Performs poorly with varying density clusters. • Inefficient in high dimensional spaces.
Gaussian Mixture Model (GMM)	• Probabilistic (soft) clustering. • Better handles overlapping clusters than K-means. • Can model elliptical cluster shapes.	• Assumes Gaussian distribution. • Requires specifying number of components. • Sensitive to initialization; may converge to local optima.

Table 2.1: Comparison of Clustering Algorithms.

5 Ensemble Clustering

Ensemble clustering [48] [49] [50] has emerged as a reliable and effective approach in the field of data clustering. It is distinguished by its ability to provide an elegant and practical solution for selecting the most appropriate clustering results, particularly in situations where the intrinsic characteristics of the dataset are unknown. By reconciling inconsistencies among varying partitions, it refines the final output and enhances overall clustering quality through the integration of evidence from multiple clustering outcomes into a unified consensus result . The ensemble clustering [51] process typically involves two primary phases: the generation phase, which produces a set of diverse partitions from the original dataset, and the consensus phase, which aggregates these partitions to derive a final, consensus-based clustering solution.

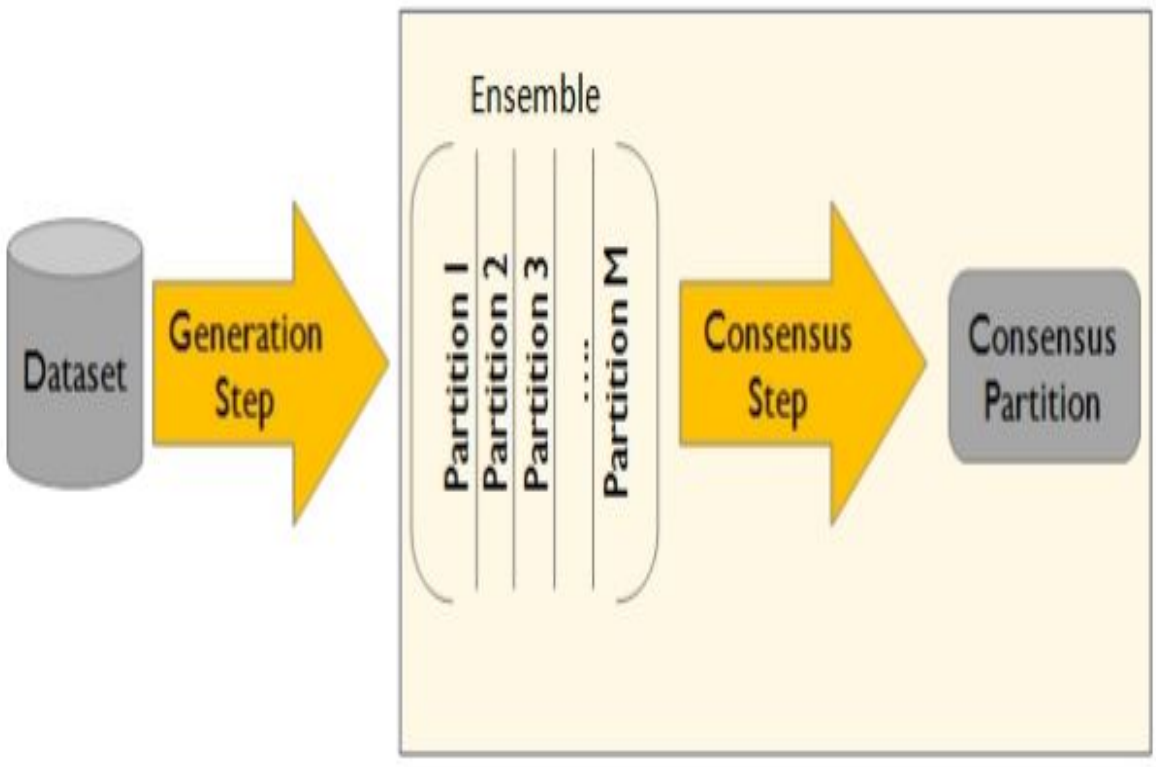


Figure 2.1: Ensemble Clustering.
[48]

5.1 Generation Process

The generation process [48] is considered the fundamental step in any ensemble clustering system. The primary objective during this phase is to produce a diverse set of partitions. The core idea is to apply clustering algorithms to the dataset multiple times (M times), ensuring that each resulting clustering is of high quality and as distinct as possible.

There are no strict constraints on how these partitions should be generated, which provides significant flexibility in employing various techniques. For instance, different clustering algorithms such as K-Means, DBSCAN, or hierarchical clustering can be used. Alternatively, the same algorithm may be run multiple times with different initializations or on random subsamples of the data. Diversity can also be introduced by preprocessing the data in different ways, such as applying dimensionality reduction or selecting different subsets of features prior to clustering [51].

This stage is crucial as it directly impacts the quality of the final results. The diversity among the partitions enables the system to capture different perspectives of the underlying structure in the data, thereby enhancing the overall accuracy of clustering. By integrating diverse and high quality partitions, the ensemble clustering system can achieve results that surpass those produced by any individual algorithm [48].

5.2 Consensus Process

The compatibility function integrates the outputs of the members to obtain a final result for the aggregation, as it represents the most important component in the aggregation system [51]. It has been classified into two main approaches: the object repetition approach and the median approach [48].

5.2.1 Co-occurrence based approach

"It firstly computes the co-occurrence of objects in the members and then determines their cluster labels to produce a consensus result. Simply, it counts the occurrence of an object in one cluster, or the occurrence of a pair of objects in the same cluster, and generates the final clustering result by a voting process among the objects."

This approach includes the following two main characteristics:

Relabeling and Voting, which addresses label inconsistency across clusterings by relabeling clusters and applying a voting mechanism to determine the final partition.

Co-associat Matrix, which constructs a matrix representing the frequency with which object pairs appear in the same cluster, serving as a similarity representation for deriving the final cluster-

ing.

5.2.2 The median partition approach

The median partition is defined as the partition that optimally maximizes similarity with all other partitions within the ensemble. It is employed to obtain a more accurate and stable consensus clustering solution.

This approach includes the following two main characteristics:

Genetic Algorithms,It seeks to determine the optimal consensus partition by maximizing a fitness function that quantifies the similarity with all individual clusterings.

Mirkin Distance, It is used to measure the degree of dissimilarity between two partitions by counting the number of object pairs that are grouped in one partition but separated in the other. It serves as an effective tool for selecting the optimal consensus partition that best approximates all individual clusterings.

6 Conclusion

In the field of machine learning, it has been proven that ensemble learning methods, which combine multiple models, perform better. Furthermore, the importance of consensus and diversity in improving accuracy has been confirmed through studies of various ensemble algorithms. These discoveries contribute to the advancement of the emotion detection field and lay the groundwork for future research aimed at further developing these technologies. The next section concentrates on an unsupervised technique used in the text based emotion detection.

CHAPTER 3

Text Based Emotion Detection

1 Introduction

With the advancement of technology and the exponential growth of the Internet, vast volumes of digitized data, including text, images, videos, and more, are now accessible through various digital social platforms. Textual data, in particular, is abundantly present across blogs, news articles, customer reviews on diverse products and services, discussion forums, recommendation systems, conversational agents, and other forms of digital interaction. These platforms provide users with the opportunity to share their thoughts and opinions freely. Consequently, they serve as rich resources for analyzing human emotions and sentiments, enabling the exploration of social trends, monitoring customer feedback to refine business strategies, and supporting consumers in their decision making processes. As a result, emotion detection from text has emerged as a prominent and impactful area of research.

This chapter explores the critical field of emotion detection from text, with a particular focus on social media platforms, notably the YouTube platform. Initially, it begins by defining text based emotion detection (ED) and differentiating it from emotion detection in long and short texts. Then it reviews the practical applications of emotion detection in short texts, including social media monitoring, customer feedback, and mental health tracking. Furthermore, the chapter discusses the challenges posed by short text data, such as sparse feature vectors, polysemy, and synonymy. It also highlights the role of text clustering techniques, such as BERT embeddings, in improving the precision of emotion detection.

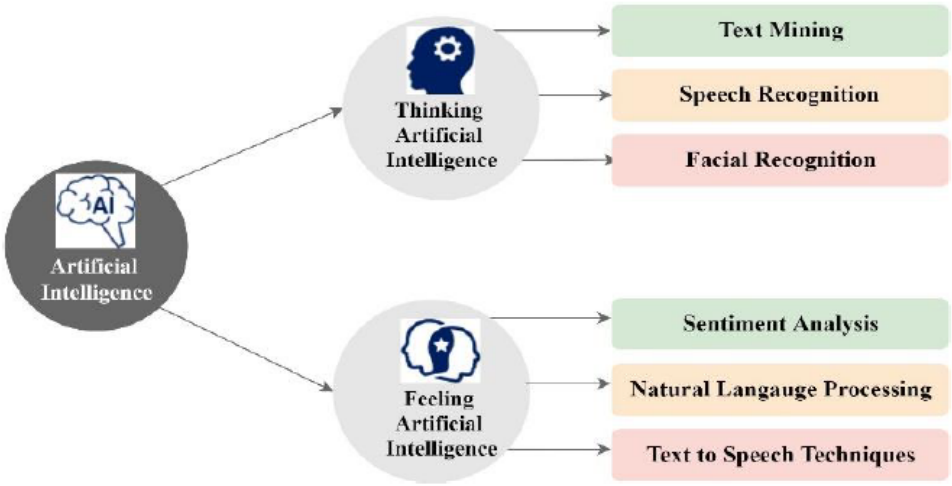


Figure 3.1: AI and the affective analysis.
[52]

2 Artificial Intelligence and the affective analysis

In recent years, artificial intelligence has expanded its domain to encompass both **Cognitive AI** (Thinking AI) and **Affective AI** (Feeling AI), Figure 3.1 illustrates the subfields within the domain of artificial intelligence. Thinking AI is primarily focused on processing information to generate new insights or make informed decisions, often utilizing unstructured data. Applications such as text mining, speech recognition, and facial detection exemplify how Thinking AI is capable of identifying patterns and extracting meaningful structures from data. Modern methodologies like machine learning and deep learning play a central role in enabling Thinking AI to effectively interpret and analyze such data.

3 Text Based Emotion Detection

Text based emotion detection [53] is a natural language processing technique aimed at identifying emotions expressed in text. This process aims to enhance human-computer interaction by enabling systems to understand emotional cues in written language.

Emotion detection in textual data has gained increasing importance across a wide range of domains, including social media analytics, customer service, learning, healthcare, and interactive entertainment. Unlike sentiment analysis, which primarily focuses on classifying text into positive, negative, or neutral categories, emotion detection aims to identify specific emotional states such

as happiness, anger, sadness, fear, or surprise. This fine grained emotional understanding provides deeper insights into user behavior and intent [54].

3.1 Emotion Detection in Long Texts

Emotion detection in long texts relies on the extraction of emotional signals from substantial written content, such as articles, reviews, and scholarly papers. These texts feature several paragraphs and varied subjects, necessitating a more profound contextual comprehension throughout the sentences. Although emotional information is rich, processing it presents computational difficulties because of its length and intricacy.

3.2 Emotion Detection in Short Texts

Emotion detection in short texts [55]emphasizes the examination of concise content like tweets, remarks, or messages, which frequently lack context and may involve slang, emojis, or abbreviations. Even though they are short, these texts can express powerful feelings. As a result, tailored models and clustering methods are utilized to improve the precision of emotion identification in these texts.

3.2.1 Word Level Emotion Detection

Word level emotion detection [56] focuses on identifying the emotional category associated with individual words within a text. This approach often relies on lexicons that map words to specific emotions such as joy, anger, and sadness. Word level methods are particularly effective in short texts, such as tweets or YouTube comments.

3.2.2 Sentence Level Emotion Detection

Sentence level emotion detection [57] aims to capture the overall emotional tone of an entire sentence rather than isolated words. This approach utilizes machine learning and deep learning models, such as Long Short-Term Memory (LSTM) networks, to understand the syntactic and semantic relationships between words.

4 Applications of Text Based Emotion Detection

Text based emotion recognition serves as a powerful tool across various fields due to the essential role of emotions in human interaction [54][15] Figure 3.2 illustrates some of the applications

of emotion detection in text. In politics, analyzing emotional content from social media enables a better understanding of public sentiment regarding security issues and policy decisions. Within the marketing domain, such analysis helps shape advertising strategies by capturing consumers' emotional responses to products and services.

In the healthcare sector, this technology contributes to the early identification of mental health disorders such as depression and anxiety through the interpretation of emotional signals in short messages and online health communities. Multiple studies have highlighted its effectiveness in detecting subtle emotional expressions within digital support environments.

On platforms like Twitter, Facebook, and YouTube, emotional analysis has been utilized to examine user reactions and comments, offering insights into behavior and enabling emotion classification for both political and social content.

Moreover, conversational agents such as chatbots have shown significant improvement in user engagement when equipped with the ability to recognize and adapt to emotional states, thereby fostering more empathetic and human-like interactions.

In customer review systems, emotion detection plays a key role in evaluating user satisfaction and improving product design based on affective feedback. From a security perspective, intelligent assistants powered by artificial intelligence have been developed to counteract social engineering attacks like phishing on social media.

Overall, the ability to interpret emotions from text represents a transformative advancement in various areas—including political analysis, mental health monitoring, commerce, and human-computer interaction underscoring its growing relevance in today's text driven digital landscape.

5 Approaches for Text based Emotion Detection

In this section, we highlight the most important approaches used for emotion detection in text:

5.1 Keyword Based Approach

The keyword-based approach is one of the oldest and simplest methods used for emotion detection in text. This method operates by identifying specific emotional words such as "happy," "angry," or "sad" within the input text, and then assigning an appropriate emotional label based on the frequency of these keywords [28]. Despite its ease of implementation and interpretation, this approach lacks the capacity to account for contextual subtleties, sarcasm, or implicit emotional cues.

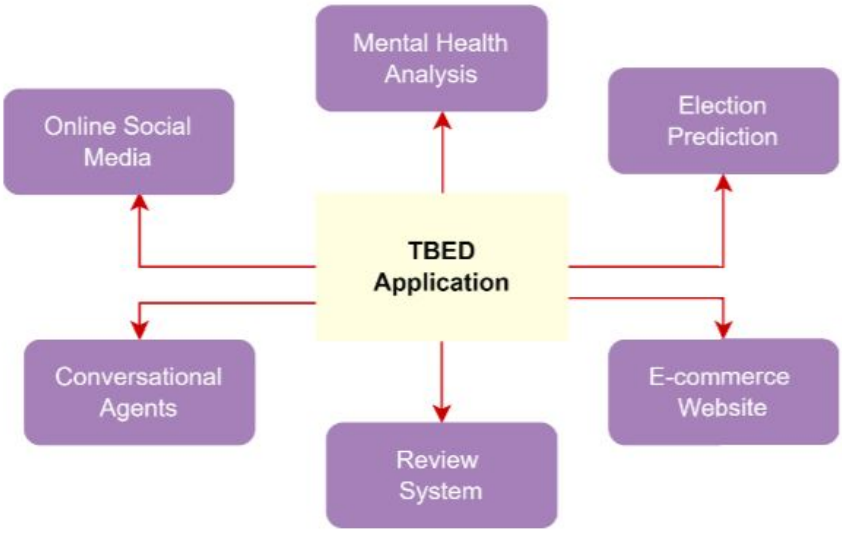


Figure 3.2: the applications of emotion detection in text.
[54]

5.2 Rule construction approach

The rule based approach is one of the oldest and simplest methods for detecting emotions in textual data. It relies on a predefined set of linguistic and logical rules, often constructed around emotional keywords or affective lexicons. These rules aim to identify emotional expressions by matching specific words or phrases within sentences, such as "happy", "sad", "angry", or "surprised"[15].

Despite the simplicity of this approach, its effectiveness largely depends on the comprehensiveness and accuracy of the emotional lexicon being used. Moreover, it faces challenges related to polysemy, contextual interpretation, and the diversity of linguistic expressions, which make it less scalable and less applicable to large and varied datasets.

5.3 Machine Learning Approach

The machine learning approach addresses the problem of emotion detection by framing it as a classification task, where algorithms learn patterns from labeled data to predict various emotion categories [15] These methods can be supervised, using annotated datasets, or unsupervised, relying on clustering techniques or latent representations.

5.4 Hybrid approach

The hybrid approach combines rule based methods with machine learning techniques, aiming to leverage the precision of linguistic rules alongside the flexibility of statistical models.Affective lexicons are employed alongside traditional classifiers or deep learning models, enhancing detection accuracy, particularly in scenarios with limited labeled data [15]. The effectiveness of this approach largely depends on the quality of integration between symbolic and statistical components, as well as the selection of the most suitable learning model.

6 Text Clustering and Emotion Detection

Text clustering is an unsupervised learning technique aimed at organizing textual data into groups based on semantic similarity. This approach plays a significant role in information retrieval, topic discovery, and content summarization. Traditional techniques such as K-means , Hierarchical Clustering and DBSCAN are commonly used; however, they often struggle with short and sparse texts due to the lack of context or structure. To address this, BERT embeddings and ensemble clustering techniques have emerged to enhance the quality and interpretability of clustering [53].

Emotion detection refers to the process of identifying emotional cues in texts, often classifying these emotions into basic categories such as joy, anger, sadness, and fear [28]. This technique has practical applications in social media monitoring, customer service, mental health tracking, and human-computer interaction.

Recent research combines clustering techniques with emotion detection to explore latent emotional patterns in unlabelled textual data. Texts are first clustered using embedding-based or probabilistic models, then the emotional distribution within each group is analyzed [58]. This approach allows for the detection of emotional themes in large sets of short and random texts, such as Twitter tweets or YouTube comments.

7 Challenges in Text Based Emotion Detection

Emotion detection in short texts, such as YouTube comments, faces various challenges that present unique difficulties. The following section will address the main challenges related to text clustering and emotion detection in this type of short text data:

7.1 Sparse Feature Vector

In clustering algorithms, each text is represented as a feature vector containing numerical values that reflect the presence or frequency of specific terms or phrases within the text. However, short texts contain very few words, which naturally results in sparse feature vectors. This sparsity poses a significant challenge for clustering algorithms.

7.2 Polysemy

Represents a significant challenge in sentiment detection from texts, as words can have multiple meanings depending on the context in which they are used, making it difficult to determine their precise meaning.

7.3 Synonymy

Identifying words with identical meanings presents a challenge in text clustering. For instance, words like 'war,' 'conflict,' 'combat,' 'armed struggle,' and 'aggression' share the same meaning. Deciding which cluster to assign such words becomes particularly difficult, especially when they appear in short texts.

8 Conclusion

In this chapter we focused on emotion detection in short texts, such as YouTube and Twitter comments, which is a challenge that requires advanced techniques to overcome issues like lack of context, polysemy, and synonymy. Text clustering techniques like BERT can play an important role in improving the accuracy of emotion detection by organizing texts into emotionally relevant groups. As research progresses, the opportunities for integrating these techniques increase, leading to the development of more accurate and effective systems for understanding emotions in short, unlabeled texts. The next section presents the detailed description of the conception and implementation of our system.

CHAPTER 4

Conception And implementation

1 Introduction

In this chapter, we present the complete description of our proposed system for unsupervised emotion detection from multilingual short text data (Arabic and English). This includes the data collection process, preprocessing, sentence embedding using pre-trained transformers (BERT and AraBERT), dimensionality reduction, and ensemble clustering. We finalize the chapter with an analysis of the clustering results and emotion interpretation using Ekman's model.

2 Conception

Emotion detection from textual data is a complex process comprising several critical stages, ranging from data collection and preprocessing to the implementation of advanced machine learning models for precise emotion classification. Figure 4.1 provides a structured visualization of the sequential steps in our methodology, delineating the data flow from initial acquisition through preprocessing and final model application. This schematic representation serves as a framework for comprehending the systematic approach employed in our research, while emphasizing the integration of diverse techniques and technologies to achieve effective emotion detection in textual data.

2.1 Collect data

In recent years, the study of emotion detection from textual data has garnered significant attention due to its wide ranging applications in areas such as sentiment analysis and customer feedback

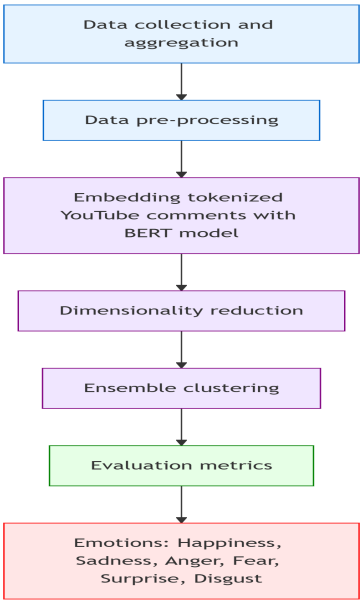


Figure 4.1: Steps of Emotion Detection.

evaluation. However, a major challenge in this field lies in collecting diverse and representative datasets necessary for effectively training emotion detection models. To overcome this issue, researchers frequently utilize various sources that provide access to comprehensive textual databases. In our study, we concentrated on one such prominent source: the YouTube API. This platform offers access to an extensive repository of textual data, covering a broad spectrum of topics and languages.

2.1.1 Youtube API

The YouTube Data API v3 [59] is a RESTful API provided by Google that allows developers to programmatically access and manipulate YouTube data. This includes operations such as retrieving video metadata (e.g., title, description, number of likes, view count, and publish date), channel information, playlists, and comments. It also enables developers to perform search queries for videos or channels.

For each comment, the API can provide detailed information such as: The text of the comment, the publish date, the number of likes on the comment, the author's display name

2.2 Data Preprocessing

Prior to emotion detection, data preprocessing is a crucial step to enhance the quality and consistency of textual input. This process comprises several tasks distributed across multiple phases, as illustrated in Figure 4.2

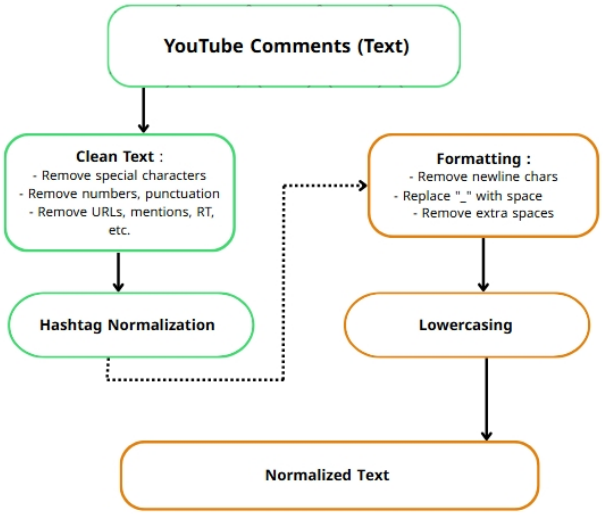


Figure 4.2: Data Preprocessing.

1. **Text Cleaning:** Initially, the text is cleaned by removing URLs, user mentions (@username), special characters, HTML tags, stock market tickers, and other extraneous elements to minimize noise.

2. **Duplicate Letter Normalization:** Repeated letters are normalized by limiting the repetition to a maximum of two consecutive characters to standardize word forms.

3. **Hashtag Processing:** Hashtags are handled by removing the # symbol while retaining the associated words, as they may contain meaningful semantic information.

4. **Formatting Adjustments:** This step involves removing newline characters, replacing underscores with spaces, and eliminating extra whitespace to enhance consistency.

5. **Lowercasing:** All text is converted to lowercase to ensure that identical words in different cases are not treated as distinct by the model.

6. **Stop Word Retention:** Although stop words (such as articles, prepositions, and conjunctions) are commonly removed to reduce noise.

7. **Number Preservation:** Numerical values are preserved during preprocessing, as they can carry significant contextual meaning (e.g., “100 deaths”, “24 hours”) that may contribute to understanding the intensity or scope of emotional content.

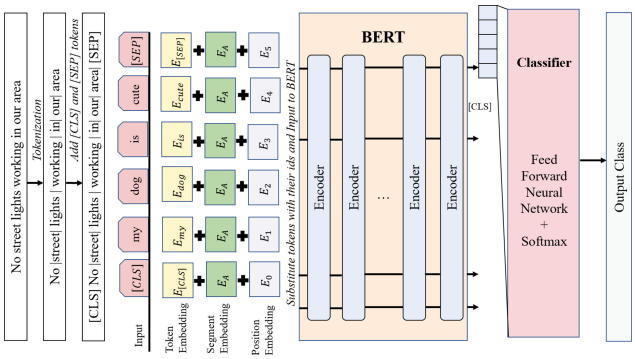


Figure 4.3: Bert-Architecture.[60]

2.3 BERT (Bidirectional Encoder Representations from Transformers)

After completing the data cleaning stage, text tokenization is performed using BERT's tokenizer. This tokenizer segments the cleaned text into subword units based on the WordPiece algorithm and converts them into token IDs that are interpretable by the BERT model [60]. These token IDs are then used as input embeddings, allowing BERT to capture the contextualized semantic representations of the text in both forward and backward directions. This bidirectional encoding is essential for tasks such as emotion detection, where understanding the full context of each word is crucial.

2.3.1 Tokenization

Tokenization constitutes a critical component of our methodology when working with BERT, serving as the essential preprocessing stage for natural language processing applications. This fundamental process transforms unstructured textual data into discrete tokens, thereby enabling BERT to comprehensively interpret and analyze linguistic subtleties. Our implementation of a specialized tokenization framework ensures optimal representation of textual content, thereby enhancing the model's capacity for sophisticated language comprehension and semantic modeling. This approach provides the necessary foundation for BERT to perform advanced natural language understanding tasks with improved accuracy and efficiency.

2.3.2 Generating Embeddings

In the embedding generation process, input text is transformed into dense vector representations referred to as embeddings which encapsulate the contextual meaning and inter-word relationships within a sentence. These representations are produced using token IDs, which serve as the input to the BERT model. The primary output of BERT consists of contextualized embeddings for each

input token, effectively encoding rich semantic and syntactic information. These embeddings are highly adaptable and can be employed in a variety of downstream natural language processing tasks.

It is worth emphasizing that when employing BERT, conventional text preprocessing steps such as stop word removal or lemmatization are typically unnecessary and may even be counterproductive. This is due to BERT's design as a pretrained language model, which is inherently capable of capturing subtle contextual nuances at both the word and subword levels without requiring extensive linguistic simplification.

2.4 Dimensionality Reduction

Dimensionality reduction [61] is a crucial step when working with high dimensional data, such as the embeddings generated by models like BERT. These embeddings often contain hundreds of features, which can introduce noise, increase computational complexity, and hinder the performance of clustering algorithms. The primary objective of dimensionality reduction is to project the data into a lower dimensional space while preserving the most relevant information. This process enhances data interpretability, improves clustering performance, and facilitates visualization. Among the various techniques available, Principal Component Analysis (PCA) is commonly used due to its effectiveness and ease of implementation. PCA transforms the original data into a set of uncorrelated variables called principal components that capture the maximum variance in the data.

2.5 Ensemble Clustering

In this project, ensemble clustering was applied as a core technique to improve the robustness and stability of unsupervised emotion detection in YouTube comments. Instead of relying on a single clustering algorithm or a fixed representation, an ensemble approach was used to combine the strengths of multiple methods. The process is implemented in three main phases: using multiple data representations, applying diverse clustering algorithms, and combining results via a co-association matrix.

2.5.1 Different Object Representations

The input comments were first encoded using pretrained transformer models. Specifically, AraBERT was used for Arabic comments and BERT for English comments. To reduce the dimensionality and enhance clustering efficiency, Principal Component Analysis (PCA) was applied to the embeddings, retaining the most informative components. This reduced representation helps in capturing meaningful semantic variations while mitigating the curse of dimensionality.

2.5.2 Different Clustering Algorithms

In this project, we employ an **ensemble clustering approach** to enhance the detection of emotions in textual data. The idea is to combine multiple clustering algorithms to benefit from their individual strengths. The process is summarized as follows:

- **Clustering Algorithms Used:**
  - **KMeans:**
    - Efficient for large datasets.
    - Captures spherical clusters.
  - **Gaussian Mixture Models (GMM):**
    - Models clusters of various shapes.
    - Supports soft clustering, allowing overlapping emotions.
  - **Agglomerative Clustering:**
    - Provides a hierarchical representation of the data.
    - Does not require a predefined number of clusters.
  - **Spectral Clustering:**
    - Detects complex, non-convex cluster structures.
  - **DBSCAN:**
    - Identifies clusters based on density.
    - Effectively detects outliers and noise points.

2.5.3 Based on Object Co-occurrence

To consolidate the results from different algorithms into a single clustering solution, a co-association matrix was constructed.

2.5.3.1 Co-association Matrix

For each clustering result, a binary matrix was generated where a value of 1 indicates that two comments were grouped in the same cluster, and 0 otherwise. These matrices were summed across all algorithms to form a final co-association matrix that captures the frequency

of object co-occurrence across the base clusterings. This matrix was then normalized and treated as a similarity matrix.

The final consensus clustering was obtained by applying agglomerative clustering to this co-association matrix, effectively grouping comments based on their consistent co-occurrence across the different base clusterings.

1. **Relabeling and Voting:** This technique involves aligning the cluster labels across different partitions using a relabeling process and then applying a majority vote to determine the final label for each object.
2. **Co-association Matrix:** A co-association matrix is built by counting how often each pair of objects appears in the same cluster across all partitions. This matrix is then used as a similarity matrix for a final clustering step, often with agglomerative clustering.

2.5.3.2 Based on Median Partition

This approach seeks to find a single partition that minimizes the overall disagreement (or distance) from all individual base clusterings.

1. Genetic Algorithms

   Genetic algorithms are employed to search for the optimal consensus partition by simulating natural evolutionary processes. They explore different combinations of cluster assignments to minimize an objective function such as pairwise disagreement.

2. Mirkin Distance

   Mirkin distance is a metric used to quantify the dissimilarity between two partitions. It can serve as an objective function for identifying a consensus partition that lies at the "median" of all base partitions.

2.6 Evaluation Metrics

Clustering is a technique used to identify similarities among data points that do not have predefined class labels. It partitions the dataset into multiple clusters, such that data points within the same cluster exhibit greater similarity to one another than to those in different clusters. Since clustering is an unsupervised learning approach, it inherently lacks direct methods for validating model accuracy. To overcome this limitation, various evaluation strategies have been developed, including internal evaluation, external evaluation, and manual assessment.

In this study, we utilize internal evaluation metrics to measure the performance of clustering algorithms. These metrics enable the assessment of cluster quality without dependence on external class labels. By examining the cohesion within clusters and the separation between clusters, internal metrics offer valuable insights into the effectiveness of the clustering methods employed. Among the internal evaluation measures considered are the following:

2.6.1 Silhouette Coefficient

The Silhouette Coefficient measures how similar a data point is to its own cluster compared to other clusters [62]. It is computed for each data point as follows:

$$SilhouetteCoefficient = \frac{b-a}{\max(a,b)} \tag{4.1}$$

Where:

- a is the mean intra cluster distance (i.e., the average distance between the point and all other points in the same cluster).
- b is the mean nearest cluster distance (i.e., the average distance between the point and all points in the nearest different cluster).

The coefficient ranges from -1 to 1:

- Values close to 1 indicate well separated and cohesive clusters.
- Values around 0 suggest overlapping clusters.
- Negative values imply incorrect cluster assignment.

2.6.2 Davies-Bouldin Index

The Davies-Bouldin Index (DBI) [63] evaluates the quality of clustering by comparing the intra cluster dispersion with the inter cluster separation. It is defined as follows:

$$DBI = \frac{1}{c}\sum_{i=1}^{c}\max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \tag{4.2}$$

Where:

- c is the number of clusters.

- σ_i and σ_j represent the average distance between each point in cluster i or j and its respective centroid (i.e., intra cluster dispersion).
- $d(c_i, c_j)$ is the distance between the centroids of clusters i and j (i.e., inter cluster distance).

2.6.3 Calinski-Harabasz Index (CHI)

The Calinski-Harabasz Index (CHI) [64] also referred to as the *Variance Ratio Criterion*, measures the clustering performance by evaluating the ratio between inter cluster and intra cluster dispersion. It is defined as:

$$CHI = \frac{trace(B_c)}{trace(W_c)} \cdot \frac{n_E - c}{c - 1} \tag{4.3}$$

Where:

- $trace(B_c)$ is the trace of the between-cluster dispersion matrix, representing the separation among different clusters.
- $trace(W_c)$ is the trace of the within-cluster dispersion matrix, indicating the compactness within each cluster.
- n_E denotes the total number of data points in the dataset.
- c is the number of clusters.

High CHI values indicate well-defined and clearly separated clusters, whereas low values suggest overlapping clusters or ineffective cluster formation.

3 Implementation

This section presents a brief overview of the programming environment and language utilized in the development of our system. In addition, it explores the detailed implementation of the system's various components, providing a thorough insight into the overall development process.

3.1 Programing Environment

In this part, we present the main tools, libraries, and programming resources employed during the development of the system. The aim is to highlight the technical environment that supported the implementation process, including the programming language, development platforms, and essential software dependencies.

3.1.1 Google Colaboratory

We utilized Google Colaboratory (Google Colab) [65], a cloud-based platform,environment for Python coding and data analysis. Its cloud accessibility, integration with Google Drive, and availability of pre-installed libraries made it convenient for handling data extraction and cleaning tasks. Although we did not rely on Colab for modeling or training, its efficient environment significantly supported the initial data preparation phase of our project.

3.1.2 Python language

Our project utilizes Python [66], a versatile and widely used high-level programming language known for its simplicity and readability. Created in 1991 by Guido van Rossum, Python supports multiple programming paradigms including procedural, object oriented, and functional programming. Being an interpreted language, Python enables rapid development and testing. Its extensive standard library and rich ecosystem of third party packages facilitate applications across various domains such as web development, data science, machine learning, automation, and scientific computing. Python's cross platform compatibility and strong community support have contributed to its global popularity and make it an ideal choice for our data analysis and machine learning tasks.

3.1.3 Scikit-learn

Scikit-learn [67] is a powerful Python library used for machine learning and data mining tasks. It provides simple and efficient tools for data preprocessing, classification, regression, clustering, and model evaluation, making it essential for implementing various algorithms in our project.

3.1.4 Random

The Python random [22] module is used to generate pseudorandom numbers for random sampling, shuffling data, and initializing stochastic algorithms within our workflow.

3.1.5 Numpy

NumPy [68] is a fundamental library for numerical computing in Python. It provides support for large, multidimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficiently.

3.1.6 Pandas

Pandas [69] is a data manipulation and analysis library built on top of NumPy. It offers data structures like DataFrames and Series, which simplify the process of cleaning, transforming, and analyzing structured data.

3.1.7 Emoji

The emoji Python library [70] allows easy handling and conversion of emojis in text data. It was used in our project to process and analyze the emotional content conveyed by emojis in comments.

3.1.8 Deap

DEAP (Distributed Evolutionary Algorithms in Python) [71] is a flexible evolutionary computation framework. It supports genetic algorithms and other optimization techniques, which can be useful for hyperparameter tuning or ensemble model optimization in machine learning projects.

3.1.9 Beautiful Soup

Beautiful Soup (bs4) [72] is a Python library designed for parsing HTML and XML documents. In our project, we used Beautiful Soup primarily for data cleaning purposes. Since some comments contained embedded HTML code, this library helped us efficiently parse

and remove HTML tags to extract clean textual content, ensuring more accurate subsequent analysis.

3.2 Comment Tokenization and Embedding

3.2.1 Hugging face

Hugging Face [73] is a leading platform and library for natural language processing (NLP). It offers a wide range of pretrained transformer models and tools that facilitate tokenization, embedding, and fine tuning for various NLP tasks. We used Hugging Face to access advanced models that helped in processing and understanding the comments efficiently.

3.2.2 Sentence Transformer

Sentence Transformer [74] is a library built on top of Hugging Face transformers, designed to generate meaningful sentence embeddings. It transforms sentences or short texts into dense vector representations that capture semantic relationships, enabling better performance in clustering and classification tasks in our project.

3.2.3 Transformers

Transformers [75] is a state of the art library by Hugging Face that provides pretrained models for natural language processing tasks such as text classification, sentiment analysis, and language modeling. It enabled us to leverage powerful models like BERT and AraBERT for feature extraction and emotion detection.

3.2.4 BERT Multilingual Base Model

To transform raw input text into rich, contextualized vector representations, we employed pretrained transformer based models that effectively capture both the semantic and syntactic nuances of words within sentences.

For English textual data, we utilized the well established **BERT-base uncased model (bert-base-uncased)**. This model has been trained on extensive English corpora and processes text in lowercase format, thereby ignoring case variations. It yields robust contextual embeddings that are well-suited for a diverse array of natural language processing tasks.

```
MODEL_NAME = "aubmindlab/bert-base-arabertv02"
preprocessor = ArabertPreprocessor(model_name=MODEL_NAME)
tokenizer = BertTokenizer.from_pretrained(MODEL_NAME)
bert_model = BertModel.from_pretrained(MODEL_NAME)
```

Figure 4.4: Bert Model.

```
def get_bert_embeddings(texts):
    # Ensure all texts are strings before tokenization
    texts = [str(text) for text in texts]
    inputs = tokenizer(texts, padding=True, truncation=True, return_tensors="pt", max_length=128)
    with torch.no_grad():
        outputs = bert_model(**inputs)
    return outputs.last_hidden_state[:, 0, :].numpy()  # CLS token representation
```

Figure 4.5: comment embedding.

In the case of Arabic text, we employed **AraBERT v2(`aubmindlab/bert-base-arabertv2`)**, a transformer-based model pre-trained on large-scale Arabic datasets. AraBERT is specifically designed to address the linguistic complexities inherent to Arabic, including its rich morphological structure and various dialectal forms. This model thus facilitates a more precise and nuanced representation of Arabic textual content, which is vital for subsequent tasks such as clustering and emotion recognition.

Both models were accessed via the Hugging Face Transformers library and operated in evaluation mode to extract embeddings without additional fine-tuning.

The function **"get_bert_embeddings"** produces contextualized embeddings for a list of input sentences by extracting the representation corresponding to the [CLS] token from the model's output. Within BERT, special tokens hold critical significance during embedding extraction:

[CLS] (Classification): Placed at the start of each input sequence, this token's final hidden state is commonly used as a holistic representation of the entire sentence. It is especially valuable for classification and clustering applications.

[SEP] (Separator): Serves to delineate boundaries between segments, particularly in sentence-pair tasks such as question answering or next sentence prediction. In single-sentence inputs, it marks the end of the sequence.

3.3 Dimensionality Reduction

The output of the BERT model had a relatively high dimensionality, which posed challenges during the clustering phase due to increased computational complexity and potential

```
pca = PCA(n_components=2)
X_reduced = pca.fit_transform(X)
print(X_reduced)
```

Figure 4.6: Dimensionality Reduction Using PCA.

```
[ 7.307227   -3.1748478 ]
[ 6.39315    -1.2040857 ]
[-6.795012   -1.6187623 ]
[-6.795012   -1.6187623 ]
[-6.795012   -1.6187623 ]
[-1.6592855   2.947046  ]
[ 2.7346034   1.7536863 ]
[-6.795012   -1.6187623 ]
[-6.795012   -1.6187623 ]
[-2.164062    1.1657342 ]
[-6.795012   -1.6187623 ]
[-6.795012   -1.6187623 ]
[ 5.894321   -0.9244215 ]
[-6.795012   -1.6187623 ]
[-0.56355226  1.6164682 ]
```

Figure 4.7: PCA Results Visualization.

noise. To address this issue, we applied Principal Component Analysis (PCA) to reduce the dimensionality of the feature space. PCA allowed us to retain the most significant variance in the data, thereby preserving essential patterns and structures while producing a more compact and manageable representation. This dimensionality reduction step enhanced the efficiency and effectiveness of the subsequent clustering process.

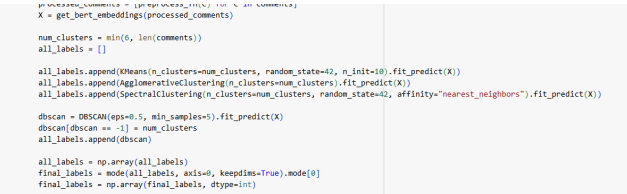
3.4 Ensemble Clustering

To enhance clustering robustness beyond simple majority voting, we applied multiple ensemble clustering strategies by varying the consensus function across iterations. Specifically, the outputs of four clustering algorithms KMeans, Agglomerative Clustering, Spectral Clustering, and DBSCAN were aggregated using different consensus mechanisms such as majority voting, weighted voting, and custom re-labeling approaches.

Each model was applied to BERT based embeddings extracted from preprocessed YouTube comments. For DBSCAN, noise points (label = -1) were reassigned to unique clusters to prevent them from disrupting the consensus.

By iteratively changing the consensus function and combining clustering results, the ensemble method improved the stability and interpretability of emotion cluster assignments,

effectively handling the diversity in comment styles.



```
X = get_bert_embeddings(processed_comments)

num_clusters = min(6, len(comments))
all_labels = []

all_labels.append(KMeans(n_clusters=num_clusters, random_state=42, n_init=10).fit_predict(X))
all_labels.append(AgglomerativeClustering(n_clusters=num_clusters).fit_predict(X))
all_labels.append(SpectralClustering(n_clusters=num_clusters, random_state=42, affinity="nearest_neighbors").fit_predict(X))

dbscan = DBSCAN(eps=0.5, min_samples=5).fit_predict(X)
dbscan[dbscan == -1] = num_clusters
all_labels.append(dbscan)

all_labels = np.array(all_labels)
final_labels = mode(all_labels, axis=0, keepdims=True).mode[0]
final_labels = np.array(final_labels, dtype=int)
```

Figure 4.8: ensemble clusetring.

3.5 Analysis of English Data

In this section, we present the analysis of the English dataset using ensemble clustering techniques. We begin with an overview of the dataset and its key characteristics, followed by a description of the preprocessing steps applied to clean and prepare the data. Finally, we present the results of our analysis, along with a discussion interpreting the key findings.

3.5.1 Dataset

Our analysis places particular emphasis on handling emojis, even though the data clearly shows that the number of emojis is significantly smaller than the number of words. Nevertheless, the expressive power of emojis is notably substantial: a single emoji can effectively convey an emotional nuance that would otherwise require multiple words. This reinforces the necessity of incorporating emojis into our emotion analysis framework, as their ability to encapsulate complex affective meanings in a compact form contributes significantly to the overall emotional interpretation of the dataset. To support this focus, we implemented a filtering mechanism to distinguish between comments that contain emojis and those that do not. This was achieved by applying a Boolean function," **"contains_emoji"**, to each comment in the dataset, as follows figure 4.9

```
df["has_emoji"] = df["Cleaned"].apply(contains_emoji)

comments_with_emoji = df[df["has_emoji"]]["Cleaned"].tolist()
comments_without_emoji = df[~df["has_emoji"]]["Cleaned"].tolist()
```

Figure 4.9: Emoji Analysis.

Figure 4.10: Script of Pre-processing Functions.

3.5.2 Preprocessing

To carry out the preprocessing steps defined in our design, we employed several Python packages. Below, we highlight some of the key libraries used, along with specific functions leveraged from each, as illustrated in Figure 4.10:

Emoji: The emoji library is a Python package designed for efficient handling and manipulation of emojis in text. It supports operations such as adding, removing, and converting emojis within strings, making it useful for emotion-aware text analysis.

RegEx: Regular Expressions offer a robust method for pattern matching and string manipulation. This tool is widely used for identifying, replacing, and extracting specific text patterns, and is essential for preprocessing tasks such as cleaning and validating textual data.

3.5.3 Results and discussion

In this section, we present the results of the ensemble clustering methods applied to the dataset. We highlight the impact of different representation levels, emoji usage, and clustering strategies using standard evaluation metrics.

3.5.4 Results of different object representations

The results presented in Table 4.1 reveal significant differences in clustering performance between comments containing emojis and those without. Across all representation levels, both word level and sentence level, the presence of emojis consistently leads to higher silhouette scores and lower Davies Bouldin indices, indicating more compact and well separated clusters.

For instance, at the sentence-level, comments with emojis achieved a Silhouette Score of 0.3132 and Davies-Bouldin Index of 1.4020, compared to 0.1637 and 1.2753 respectively for without emojis. While the Davies Bouldin Index is slightly better without emojis, the

Silhouette score (a more intuitive measure of cluster cohesion and separation) is substantially better with emojis. This supports the hypothesis that emojis enrich the emotional content of text, helping the clustering algorithm to more accurately group semantically similar emotional expressions.

2. Comparison Across Representation Levels

Another important observation is the improvement in clustering performance when moving from word-level to sentence-level representations. Sentence-level embeddings, which capture the holistic meaning of a comment rather than isolated keywords, result in more coherent emotional groupings.

This is reflected in the Calinski-Harabasz Index, which jumps from 33.7582 (word-level with emojis) to 107.9662 (sentence-level with emojis). This index favors compact clusters that are well-separated from one another, and such a jump suggests that sentence-level embeddings offer a better structural foundation for emotion clustering.

Table 4.1: Clustering Evaluation Metrics for Word-level and Sentence-level Representations (With and Without Emoji).

Representation Level	Emoji Usage	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Word-level	With Emoji	0.2100	33.7582	2.2829
	Without Emoji	0.1125	58.5225	2.7659
Sentence-level	With Emoji	0.3132	107.9662	1.4020
	Without Emoji	0.1637	504.4816	1.2753

3.5.5 Results of Co-occurrence based approach

Table 4.2 compares two different ensemble techniques: voting based and co-association matrixapproaches. Both techniques show higher clustering performance with emojis.

The Co-association method yields slightly better results overall than the voting method for with emoji data, especially in the Silhouette Score (0.1883 vs 0.1526). This suggests that merging clustering outputs through shared cluster memberships (as in co-association) captures the underlying structure more effectively than simple majority voting.

On the other hand, both methods perform poorly on data without emojis, with silhouette scores close to zero (or negative), reflecting the challenge of detecting emotional clusters when the expressive cues (i.e., emojis) are absent.

Table 4.2: Comparison of Clustering Evaluation Metrics: Voting and Co-association matrix.

Method	Emoji Usage	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
Voting	With Emoji	0.1526	18.5129	2.1426
	Without Emoji	0.0271	32.3890	3.5419
Co-association	With Emoji	0.1883	22.6929	2.5532
	Without Emoji	0.0043	53.7104	2.9119

3.5.6 Results of The median partition approach

Table 4.3 highlights the results of median-based consensus clustering methods, specifically the Mirkin metric and MM-GA (Genetic Algorithm based approach).

Here again, data with emojis consistently shows better clustering quality across all metrics. The MM-GA approach slightly outperforms the Mirkin metric across all scores:

For instance, with emojis: Silhouette (0.1239 vs 0.0987), Davies-Bouldin (4.6778 vs 5.7308). However, the overall scores remain relatively low, especially for without emoji data, where silhouette scores are negative, indicating poor clustering quality and possibly the presence of noise or overlapping clusters.

This highlights that median partitioning methods are sensitive to noise and may not capture complex emotional structures well unless enriched representations (such as those including emojis or sentence embeddings) are used.

Table 4.3: Comparison of Clustering Evaluation Metrics: Mirkin and MM-GA Approaches.

Method	Emoji Usage	Silhouette	Calinski-Harabasz	Davies-Bouldin
Mirkin	With Emoji	0.0987	11.1947	5.7308
	Without Emoji	-0.0141	17.4453	9.6689
MM-GA	With Emoji	0.1239	12.1733	4.6778
	Without Emoji	-0.0204	18.7786	9.2973

3.6 Analysis of Arabic Data

This section presents an analysis of Arabic YouTube comments using ensemble clustering techniques for emotion detection. It includes a description of the dataset, preprocessing meth-

ods, clustering results, and a comparative evaluation across different text representations and ensemble approaches.

3.6.1 Dataset

The Arabic dataset comprises user generated comments extracted from YouTube videos related to the Palestinian cause. These comments are written in both Modern Standard Arabic and regional dialects. For analytical purposes, the dataset was divided into two subsets: comments containing emojis and those without emojis. This distinction was made to assess the influence of emoji presence on the emotional clarity and quality of clustering. Each comment was further transformed into two forms of representation word-level and sentence-level to enable a comprehensive multi-level analysis of emotion laden content.

3.6.2 Preprocessing

A set of preprocessing steps was applied to normalize and clean the Arabic text. These included:

- Removing diacritics and special characters.
- Eliminating repeated letters, hyperlinks, and stopwords.
- Separating comments based on the presence or absence of emojis.

Following preprocessing, the comments were converted into numerical representations using two abstraction levels:

- **Word-level representation:** where individual words are represented using pretrained word embeddings.
- **Sentence-level representation:** where the entire comment is encoded using contextual models, such as transformer-based sentence embeddings.

This dual representation helped in assessing how abstraction level impacts the structure and cohesion of the generated clusters.

```
text = strip_tashkeel(text)
text = strip_tatweel(text)
return text.strip().lower()
```

Figure 4.11: Pre-processing additionale Functions.

3.6.3 Discussion of the results

The clustering performance was evaluated using three standard metrics: Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index.

The best performance was recorded using sentence-level representation with the presence of emojis, achieving the following:

- **Silhouette Score** = 0.3367, indicating well-defined and cohesive clusters.
- **Calinski-Harabasz Index** = 124.91, reflecting strong separation between clusters.
- **Davies-Bouldin Index** = 1.8239, the lowest among all scenarios, suggesting minimal overlap between clusters.

It is worth noting that removing emojis from the sentence-level representation still yielded a strong Calinski-Harabasz score (208.52); however, both the Silhouette and Davies-Bouldin scores showed a decline in performance, confirming that the presence of emojis enhances the emotional structure of the clusters.

On the other hand, word-level representations exhibited weaker performance across all metrics, especially in the absence of emojis, where the lowest Silhouette Score (0.1169) and the highest Davies-Bouldin Index (2.2256) were observed. This indicates poorly structured and incohesive clusters.

3.6.4 Results of different object representations

The comparative evaluation between word-level and sentence-level representations, in both the presence and absence of emojis, revealed that sentence-level representation with emojis consistently outperforms other configurations.

These findings support the hypothesis that a higher level of abstraction, combined with emotional cues such as emojis, contributes to improved clustering quality.

Table 4.4: Clustering Evaluation Metrics for Word-level and Sentence-level Representations (With and Without Emoji).

Emoji Usage	Representation Level	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
With Emoji	Word-level	0.3126	91.0094	2.2704
	Sentence-level	0.3367	124.9168	1.8239
Without Emoji	Word-level	0.1169	82.6422	2.2256
	Sentence-level	0.1694	208.5258	1.9197

3.6.5 Results of Co-occurrence based approach

The evaluation of co-occurrence based ensemble clustering methods revealed notable differences in performance between the Voting strategy and the Co-association Matrix approach. The Co-association Matrix method, particularly when applied to comments containing emojis, yielded the most favorable results across all clustering evaluation metrics. It achieved higher cluster cohesion and separation, as indicated by its superior Silhouette and Calinski-Harabasz scores, along with a relatively low Davies-Bouldin score.

In contrast, the Voting method showed comparatively weaker performance, especially when emojis were included. These findings underscore the effectiveness of the Co-association Matrix technique in capturing emotional nuances in text data. The presence of emojis further enhanced its performance, suggesting that emojis contribute valuable affective cues that improve the quality of emotion-based clustering.

Table 4.5: Clustering Evaluation Metrics for Voting and Co-association Matrix Methods

Method	Category	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
Voting	With Emoji	0.0743	33.2787	2.3990
	Without Emoji	0.0729	61.0021	2.2310
Co-association Matrix	With Emoji	0.3130	92.3928	2.1606
	Without Emoji	0.0766	75.7016	1.6831

3.6.6 Results of The median partition approach

This section compares the performance of two median partition techniques Mirkin and Multi-Metric Genetic Algorithm (MM-GA) in clustering Arabic YouTube comments with and without emojis. The MM-GA approach outperformed the traditional Mirkin method in most evaluation metrics, especially when emojis were present. It achieved higher Silhouette

and Calinski-Harabasz scores, alongside a slightly lower Davies-Bouldin score, indicating more coherent and well separated clusters.

In contrast, the Mirkin method showed notably poor performance, with negative Silhouette scores and higher Davies-Bouldin values, particularly in the presence of emojis. These results suggest that MM-GA is more robust in capturing emotional structures in text data, benefiting further from the presence of emojis as affective indicators.

Overall, the results demonstrate the superiority of MM-GA over Mirkin in median partition based ensemble clustering, supporting its suitability for emotion detection tasks involving noisy and diverse datasets such as Arabic YouTube comments.

Table 4.6: Clustering evaluation comparison between Mirkin and MM_GA methods

Method	Category	Silhouette Score	Calinski-Harabasz Score	Davies-Bouldin Score
Mirkin	With Emoji	-0.0596	29.22	4.5781
	Without Emoji	-0.0086	36.68	4.5237
MM_GA	With Emoji	0.0755	64.69	4.2528
	Without Emoji	-0.0406	48.22	6.3287

The best results, according to the evaluation metrics, were obtained using both the English and Arabic datasets, as illustrated in the graphical representation below:

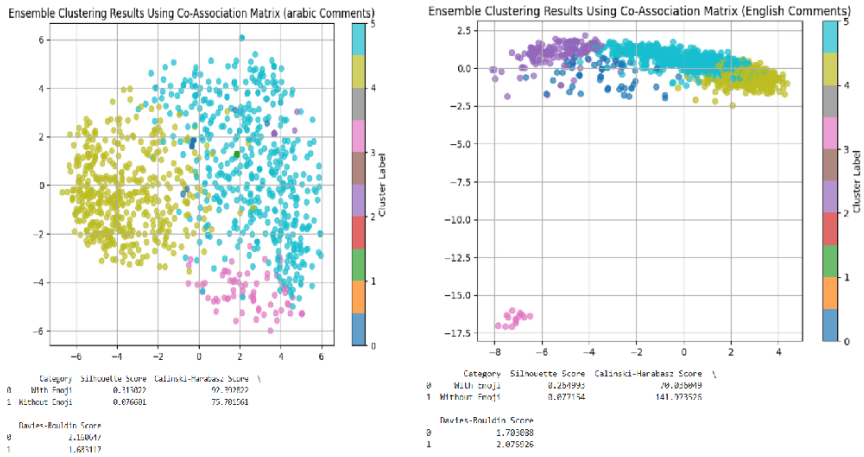


Figure 4.12: best result.

4 Conclusion

In this chapter, we explored our methodology for detecting emotions from YouTube comments related to the Gaza War, using both English and Arabic data. We aimed to improve emotion clustering by enhancing ensemble clustering techniques. The chapter covered the main steps of our approach, including data processing and the use of BERT based embeddings, followed by the presentation and discussion of results. In the next chapter, we will present our final conclusions and perspectives.

General Conclusion

This study presented a comprehensive and innovative framework for unsupervised emotion detection in multilingual YouTube comments related to the Gaza war, focusing on English and Arabic texts. By leveraging ensemble clustering techniques and varying levels of textual representation, at both word and sentence levels, with and without emojis, the proposed approach demonstrated notable improvements in clustering quality and emotional insight.

Our findings reveal that sentence-level representations enriched with emojis yielded the highest performance across all evaluation metrics, including the Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. Emojis, in particular, proved to be strong affective cues, especially in Arabic content, where textual ambiguity is more prevalent. Ensemble methods such as co-association matrices and MM-GA (Multi-Member Genetic Algorithms) outperformed traditional consensus strategies, offering better cohesion and separation of emotional clusters.

Transformer based models like AraBERT and BERT significantly enhanced the semantic understanding of short comments, especially in politically sensitive contexts like the Gaza war. The unsupervised nature of this approach also overcame the frequent limitation of lacking labeled datasets in such domains, demonstrating the viability and necessity of unsupervised methods in affective computing.

Despite the successes, challenges remain, particularly in handling dialectal variability, noise, and the absence of emotional markers such as emojis. These limitations open opportunities for future work, including adaptive weighting schemes, more robust ensemble generation strategies, integration of multimodal data (e.g., images or video cues), and advanced deep learning techniques such as subspace clustering and optimization based consensus models.

Beyond the technical contribution, this research provided insights into the emotional dy-

namics of digital interactions during conflict, contributing to fields such as digital activism, public opinion analysis, mental health surveillance, and media moderation.

In conclusion, this work successfully achieved its objective by proposing a practical, flexible, and robust framework for emotion detection in multilingual and politically charged social media content. It lays a strong foundation for future research aiming to enhance the interpretability and applicability of unsupervised emotion analysis across platforms, languages, and global crises.

Bibliography

[1] Paul Ekman. ``Basic Emotions." In: *Handbook of Cognition and Emotion*. Ed. by Tim Dalgleish and Mick J. Power. 1st ed. Wiley, Feb. 25, 1999, pp. 45–60. ISBN: 978-0-471-97836-7 978-0-470-01349-6. DOI: 10.1002/0470013494.ch3. URL: https://onlinelibrary.wiley.com/doi/10.1002/0470013494.ch3 (visited on 02/03/2025).

[2] Daniel Goleman. *Emotional Intelligence: Why It Can Matter More Than IQ*. Google-Books-ID: Lq18kigs7m0C. A&C Black, July 20, 2009. 351 pp. ISBN: 978-1-4088-0620-3.

[3] *Universal Emotions*. Paul Ekman Group. URL: https://www.paulekman.com/universal-emotions/ (visited on 01/29/2025).

[4] Robert W. Levenson. ``The Intrapersonal Functions of Emotion." In: *Cognition & Emotion* 13.5 (Sept. 1999), pp. 481–504. ISSN: 0269-9931, 1464-0600. DOI: 10.1080/026999399379159. URL: http://www.tandfonline.com/doi/abs/10.1080/026999399379159 (visited on 01/29/2025).

[5] Roy F. Baumeister et al. ``How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather Than Direct Causation." In: *Personality and Social Psychology Review* 11.2 (May 2007), pp. 167–203. ISSN: 1088-8683, 1532-7957. DOI: 10.1177/1088868307301033. URL: https://journals.sagepub.com/doi/10.1177/1088868307301033 (visited on 01/29/2025).

[6] Lauren J. Chapman and David J. Mckenzie. ``Chapter 2 Behavioral Responses and Ecological Consequences." In: *Hypoxia*. Ed. by Jeffrey G. Richards, Anthony P. Farrell, and Colin J. Brauner. Vol. 27. Fish Physiology. Academic Press, 2009, pp. 25–77. DOI:

https://doi.org/10.1016/S1546-5098(08)00002-2. URL: https://www.sciencedirect.com/science/article/pii/S1546509808000022.

[7] James J. Gross and Robert W. Levenson. ``Hiding feelings: The acute effects of inhibiting negative and positive emotion." In: *Journal of Abnormal Psychology* 106.1 (1997). Place: US Publisher: American Psychological Association, pp. 95–103. ISSN: 1939-1846(Electronic),0021-843X(Print). DOI: 10.1037/0021-843X.106.1.95.

[8] *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/complex-emotion (visited on 01/29/2025).

[9] *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/emotion (visited on 01/29/2025).

[10] *APA Dictionary of Psychology*. URL: https://dictionary.apa.org/mood (visited on 01/29/2025).

[11] Andrew Ortony and Gerald Clore. ``Can an appraisal model be compatible with psychological constructionism." In: *The psychological construction of emotion* (2015), pp. 305–333.

[12] James A. Russell. ``A circumplex model of affect." en. In: *Journal of Personality and Social Psychology* 39.6 (Dec. 1980), pp. 1161–1178. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/h0077714. URL: https://doi.apa.org/doi/10.1037/h0077714 (visited on 01/29/2025).

[13] Andrea Bolioli et al. ``A complementary account to emotion extraction and classification in cultural heritage based on the Plutchik's theory." en. In: *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. Barcelona Spain: ACM, July 2022, pp. 374–382. ISBN: 978-1-4503-9232-7. DOI: 10.1145/3511047.3537659. URL: https://dl.acm.org/doi/10.1145/3511047.3537659 (visited on 01/30/2025).

[14] Iris Bakker et al. ``Pleasure, Arousal, Dominance: Mehrabian and Russell revisited." en. In: *Current Psychology* 33.3 (Sept. 2014), pp. 405–421. ISSN: 1046-1310, 1936-4733. DOI: 10.1007/s12144-014-9219-4. URL: http://link.springer.com/10.1007/s12144-014-9219-4 (visited on 03/04/2025).

[15] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo□Mensah. ``Text□based emotion detection: Advances, challenges, and opportunities." en. In: *Engineering Reports* 2.7 (July 2020), e12189. ISSN: 2577-8196, 2577-8196. DOI: 10.1002/eng2.12189. URL: https://onlinelibrary.wiley.com/doi/10.1002/eng2.12189 (visited on 01/30/2025).

[16] Md Mahbubur Rahman and Shaila Shova. ``Emotion Detection From Social Media Posts." en. In: (Feb. 2023). arXiv:2302.05610 [cs]. DOI: 10.48550/arXiv.2302.05610. URL: http://arxiv.org/abs/2302.05610 (visited on 05/28/2025).

[17] Bharat Gaind, Varun Syal, and Sneha Padgalwar. ``Emotion Detection and Analysis on Social Media." In: arXiv:1901.08458 (June 12, 2019). DOI: 10.48550/arXiv.1901.08458. arXiv: 1901.08458[cs]. URL: http://arxiv.org/abs/1901.08458 (visited on 06/02/2025).

[18] Computer Science and Engineering Department, Guru Gobind Singh Indraprastha University, New Delhi, India et al. ``Emotion Recognition and Detection Methods: A Comprehensive Survey." en. In: *Journal of Artificial Intelligence and Systems* 2.1 (2020), pp. 53–79. ISSN: 26422859. DOI: 10.33969/AIS.2020.21005. URL: https://iecscience.org/jpapers/46 (visited on 03/04/2025).

[19] Rasmus Schmøkel and Michael Bossetta. ``FBAdLibrarian and Pykognition: open science tools for the collection and emotion detection of images in Facebook political ads with computer vision." en. In: *Journal of Information Technology & Politics* 19.1 (Jan. 2022), pp. 118–128. ISSN: 1933-1681, 1933-169X. DOI: 10.1080/19331681.2021.1928579. URL: https://www.tandfonline.com/doi/full/10.1080/19331681.2021.1928579 (visited on 05/29/2025).

[20] Advanced Technologies Institute Bucharest, Romania et al. ``RED: A Novel Dataset for Romanian Emotion Detection from Tweets." en. In: *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*. INCOMA Ltd. Shoumen, BULGARIA, 2021, pp. 291–300. ISBN: 978-954-452-072-4. DOI: 10.26615/978-954-452-072-4_034. URL: https://acl-bg.org/proceedings/2021/RANLP%202021/pdf/2021.ranlp-1.34.pdf (visited on 05/29/2025).

[21] Vikas Maheshkar and Sachin Kumar Sarin. *Review and Analysis of Emotion Detection from Tweets Using Twitter Datasets*. 2022.

[22] Abeera V P, Dr. Sachin Kumar, and Dr. Soman K P. ``Social Media Data Analysis for Malayalam YouTube Comments: Sentiment Analysis and Emotion Detection using ML and DL Models." In: (Sept. 2023). Ed. by Bharathi R. Chakravarthi et al., pp. 43–51. URL: https://aclanthology.org/2023.dravidianlangtech-1.6/.

[23] Md Mahbubur Rahman and Shaila Shova. *Emotion Detection From Social Media Posts*. Feb. 11, 2023. DOI: 10.48550/arXiv.2302.05610. arXiv: 2302.05610[cs]. URL: http://arxiv.org/abs/2302.05610 (visited on 06/02/2025).

[24] Hanzun Zhang. ``A Study of Human Emotion Analysis Based on Social Media." In: *Proceedings of the 2023 2nd International Conference on Social Sciences and Humanities and Arts (SSHA 2023)*. Ed. by Mohd Fauzi Bin Sedon et al. Vol. 752. Series Title: Advances in Social Science, Education and Humanities Research. Paris: Atlantis Press SARL, 2023, pp. 174–180. ISBN: 978-2-38476-061-9 978-2-38476-062-6. DOI: 10.2991/978-2-38476-062-6_23. URL: https://www.atlantis-press.com/doi/10.2991/978-2-38476-062-6_23 (visited on 06/02/2025).

[25] Masengu Reason, Tsikada Charles, and Garwi Jabulani. *AI and Machine Learning Applications in Supply Chains and Marketing*. Google-Books-ID: vFwrEQAAQBAJ. IGI Global, Oct. 18, 2024. 488 pp. ISBN: 979-8-3693-6762-9.

[26] *Understanding Environmental Posts Sentiment and Emotion Analysis of Social Media Data*.

[27] World Health Organization. *The World Health Report 2001: Mental Health : New Understanding, New Hope*. World Health Organization, 2001. 204 pp. ISBN: 978-92-4-156201-0.

[28] GR Reddy et al. ``Emotion detection from text and analysis of future work: A survey." In: *Rivista Italiana di Filosofia Analitica Junior* 14.1 (2023), pp. 59–73.

[29] Belal Aldabbour et al. ``Psychological impacts of the Gaza war on Palestinian young adults: a cross-sectional study of depression, anxiety, stress, and PTSD symptoms." In: *BMC Psychology* 12.1 (Nov. 26, 2024), p. 696. ISSN: 2050-7283. DOI: 10.1186/s40359-024-02188-5. URL: https://bmcpsychology.biomedcentral.com/articles/10.1186/s40359-024-02188-5 (visited on 01/30/2025).

[30] Simon Hofmann et al. *Hate Speech and Sentiment of YouTube Video Comments From Public and Private Sources Covering the Israel-Palestine Conflict*. Mar. 3, 2025. DOI: 10.48550/arXiv.2503.10648. arXiv: 2503.10648[cs]. URL: http://arxiv.org/abs/2503.10648 (visited on 05/08/2025).

[31] Ashagrew Liyih et al. ``Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning." In: *Scientific Reports* 14.1 (June 13, 2024), p. 13647. ISSN: 2045-2322. DOI: 10.1038/s41598-024-63367-3. URL: https://www.nature.com/articles/s41598-024-63367-3 (visited on 05/08/2025).

[32] Ayşen Temel EgiNli and Neslihan Özmelek Taş. ``Emotions on Social Media: A Sentiment Analysis Approach Based on Twitter (X) Data on the Russian-Ukraine War." In: *International Journal of Social Inquiry* 16.2 (Dec. 31, 2023), pp. 445–459. ISSN: 1307-8364. DOI: 10.37093/ijsi.1336016. URL: http://dergipark.org.tr/en/doi/10.37093/ijsi.1336016 (visited on 05/08/2025).

[33] Leo Ramos and Oscar Chang. ``Sentiment Analysis of Russia-Ukraine Conflict Tweets Using RoBERTa." In: *Uniciencia* 37.1 (June 1, 2023), pp. 1–11. ISSN: 2215-3470. DOI: 10.15359/ru.37-1.23. URL: https://www.revistas.una.ac.cr/index.php/uniciencia/article/view/17665 (visited on 05/08/2025).

[34] Sara Nabhani et al. ``Integrating Argumentation Features for Enhanced Propaganda Detection in Arabic Narratives on the Israeli War on Gaza." In: *Proceedings of the first International Workshop on Nakba Narratives as Language Resources*. Ed. by Mustafa Jarrar, Habash Habash, and Mo El-Haj. Abu Dhabi: Association for Computational Linguistics, Jan. 2025, pp. 127–149. URL: https://aclanthology.org/2025.nakbanlp-1.14/.

[35] YOUNES ALLAL. ``Comparison of NLP Techniques for sentiment analysis on social data (application case: war in GAZA)." In: (2024).

[36] Lynnette Hui Xian Ng, Adrian Xuan Wei Lim, and Roy Ka-Wei Lee. ``Love-Hate Dataset: A Multi-Modal Multi-Platform Dataset Depicting Emotions in the 2023 Israel-Hamas War." In: *Companion Proceedings of the ACM Web Conference 2024*. WWW '24: The ACM Web Conference 2024. Singapore Singapore: ACM, May 13, 2024, pp. 1807–1815. ISBN: 979-8-4007-0172-6. DOI: 10.1145/3589335.3651966. URL: https://dl.acm.org/doi/10.1145/3589335.3651966 (visited on 05/31/2025).

[37] *What Is Machine Learning (ML)? | IBM*. Sept. 22, 2021. URL: https://www.ibm.com/think/topics/machine-learning (visited on 02/01/2025).

[38] Paolo Trunfio. *Service-Oriented Distributed Knowledge Discovery*. Vol. 20121229. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, Oct. 15, 2012. ISBN: 978-1-4398-7531-5 978-1-4398-7533-9. DOI: 10.

1201/b12990. URL: http://www.crcnetbase.com/doi/book/10.1201/b12990 (visited on 05/31/2025).

[39] Iqbal H. Sarker et al. ``Cybersecurity data science: an overview from machine learning perspective." In: *Journal of Big Data* 7.1 (Dec. 2020), p. 41. ISSN: 2196-1115. DOI: 10.1186/s40537-020-00318-5. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00318-5 (visited on 05/31/2025).

[40] Taeho Jo. *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. Cham: Springer International Publishing, 2021. ISBN: 978-3-030-65899-1 978-3-030-65900-4. DOI: 10.1007/978-3-030-65900-4. URL: http://link.springer.com/10.1007/978-3-030-65900-4 (visited on 05/31/2025).

[41] Leslie P. Kaelbling, Michael L. Littman, and Andrew W. Moore. ``Reinforcement Learning: A Survey." In: *Journal of Artificial Intelligence Research* 4 (1996), pp. 237–285. DOI: 10.1613/JAIR.301.

[42] Laurens Van Der Maaten, Eric O Postma, H Jaap Van Den Herik, et al. ``Dimensionality reduction: A comparative review." In: *Journal of machine learning research* 10.66-71 (2009), p. 13.

[43] A. K. Jain, M. N. Murty, and P. J. Flynn. ``Data clustering: a review." In: *ACM Comput. Surv.* 31.3 (Sept. 1999), pp. 264–323. ISSN: 0360-0300. DOI: 10.1145/331499.331504. URL: https://doi.org/10.1145/331499.331504.

[44] Santosh Kumar Uppada. ``Centroid based clustering algorithms—A clarion study." In: *International Journal of Computer Science and Information Technologies* 5.6 (2014), pp. 7309–7313.

[45] Sarah P. Preheim et al. ``Distribution-Based Clustering: Using Ecology To Refine the Operational Taxonomic Unit." In: *Applied and Environmental Microbiology* 79.21 (Nov. 2013), pp. 6593–6603. ISSN: 0099-2240, 1098-5336. DOI: 10.1128/AEM.00342-13. URL: https://journals.asm.org/doi/10.1128/AEM.00342-13 (visited on 05/31/2025).

[46] Bernd Fischer, Volker Roth, and Joachim Buhmann. ``Clustering with the Connectivity Kernel." In: 16 (2003). Ed. by S. Thrun, L. Saul, and B. Schölkopf. URL: https://proceedings.neurips.cc/paper_files/paper/2003/file/cc0991344c3d760ae42259064Paper.pdf.

[47] Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. ``A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In: ().

[48] Daniel Duarte Abdala. ``Ensemble and Constrained Clustering with Applications." In: (2010).

[49] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. *Human emotion recognition: Review of sensors and methods*. 2020. DOI: 10.3390/s20030592.

[50] Nadjia Khatir. *Les techniques de clustering dédiées aux données multimédia*. Thèse de Doctorat en Informatique, spécialité STIC. Soutenue le 9 juillet 2019. Oran, Algérie, 2019. URL: https://dspace.univ-oran1.dz/handle/123456789/661.

[51] Sandro Vega-Pons and José Ruiz-Shulcloper. *A SURVEY OF CLUSTERING ENSEMBLE ALGORITHMS*. May 2011. DOI: 10.1142/S0218001411008683. URL: https://www.worldscientific.com/doi/abs/10.1142/S0218001411008683 (visited on 06/02/2025).

[52] Mingqi Gao et al. ``Deep learning for video object segmentation: a review." en. In: *Artificial Intelligence Review* 56.1 (Jan. 2023), pp. 457–531. ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-022-10176-7. URL: https://link.springer.com/10.1007/s10462-022-10176-7 (visited on 05/29/2025).

[53] Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. ``Transformer models for text-based emotion detection: a review of BERT-based approaches." en. In: *Artificial Intelligence Review* 54.8 (Dec. 2021), pp. 5789–5829. ISSN: 0269-2821, 1573-7462. DOI: 10.1007/s10462-021-09958-2. URL: https://link.springer.com/10.1007/s10462-021-09958-2 (visited on 05/29/2025).

[54] Abdullah Al Maruf et al. ``Challenges and Opportunities of Text-Based Emotion Detection: A Survey." en. In: *IEEE Access* 12 (2024), pp. 18416–18450. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3356357. URL: https://ieeexplore.ieee.org/document/10409495/ (visited on 05/30/2025).

[55] Shuai Yuan, Huan Huang, and Linjing Wu. ``Use of Word Clustering to Improve Emotion Recognition from Short Text." en. In: *Journal of Computing Science and Engineering* 10.4 (Dec. 2016), pp. 103–110. ISSN: 1976-4677. DOI: 10.5626/JCSE.2016.10.4.103. URL: http://koreascience.or.kr/journal/view.jsp?kj=E1EIKI&py=2016&vnc=v10n4&sp=103 (visited on 05/29/2025).

[56] Saif M. Mohammad and Peter D. Turney. *Crowdsourcing a Word-Emotion Association Lexicon*. en. arXiv:1308.6297 [cs]. Aug. 2013. DOI: 10.48550/arXiv.1308.6297. URL: http://arxiv.org/abs/1308.6297 (visited on 05/29/2025).

[57] Muhammad Zubair Asghar et al. ``Sentence-Level Emotion Detection Framework Using Rule-Based Classification." en. In: *Cognitive Computation* 9.6 (Dec. 2017), pp. 868–894. ISSN: 1866-9956, 1866-9964. DOI: 10.1007/s12559-017-9503-3. URL: http://link.springer.com/10.1007/s12559-017-9503-3 (visited on 05/29/2025).

[58] Bommisetty Durga Jasvitha et al. ``Emotion Detection from Text Using ML Framework." en. In: *2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE)*. Shivamogga, India: IEEE, May 2024, pp. 1–6. ISBN: 979-8-3503-7156-7. DOI: 10.1109/AMATHE61652.2024.10582116. URL: https://ieeexplore.ieee.org/document/10582116/ (visited on 05/30/2025).

[59] *Détails de l'API/du service – API et services – My First Project – Console Google Cloud*. URL: https://console.cloud.google.com/apis/api/youtube.googleapis.com/metrics?inv=1&invt=Aby5UA&project=automatic-opus-460910-k3 (visited on 05/31/2025).

[60] Amir Hossein Oliaee et al. ``Using Bidirectional Encoder Representations from Transformers (BERT) to classify traffic crash severity types." In: *Natural Language Processing Journal* 3 (June 2023), p. 100007. ISSN: 29497191. DOI: 10.1016/j.nlp.2023.100007. URL: https://linkinghub.elsevier.com/retrieve/pii/S2949719123000043 (visited on 05/31/2025).

[61] C. O. S. Sorzano, J. Vargas, and A. Pascual. ``A survey of dimensionality reduction techniques." In: *arXiv preprint arXiv:1403.2877* (2014).

[62] Duy-Tai Dinh, Tsutomu Fujinami, and Van-Nam Huynh. *Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient*. Ed. by Jian Chen et al. Series Title: Communications in Computer and Information Science. Singapore, 2019. DOI: 10.1007/978-981-15-1209-4_1. URL: https://link.springer.com/10.1007/978-981-15-1209-4_1 (visited on 06/02/2025).

[63] Frédéric Ros, Rabia Riad, and Serge Guillaume. *PDBI: A partitioning Davies-Bouldin index for clustering evaluation*. Apr. 2023. DOI: 10.1016/j.neucom.2023.01.043.

URL: https://linkinghub.elsevier.com/retrieve/pii/S0925231223000528 (visited on 06/02/2025).

[64] Lim Eng Aik, Tan Wee Choon, and Mohd Syafarudy Abu. *K-means Algorithm Based on Flower Pollination Algorithm and Calinski-Harabasz Index*. Nov. 1, 2023. DOI: 10.1088/1742-6596/2643/1/012019. URL: https://iopscience.iop.org/article/10.1088/1742-6596/2643/1/012019 (visited on 06/02/2025).

[65] *Google Colab*. URL: https://colab.research.google.com/ (visited on 05/31/2025).

[66] *Welcome to Python.org*. Python.org. May 26, 2025. URL: https://www.python.org/ (visited on 06/01/2025).

[67] *scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation*. URL: https://scikit-learn.org/stable/ (visited on 06/01/2025).

[68] *NumPy*. URL: https://numpy.org/ (visited on 06/01/2025).

[69] *pandas - Python Data Analysis Library*. URL: https://pandas.pydata.org/ (visited on 06/01/2025).

[70] *re — Regular expression operations*. Python documentation. URL: https://docs.python.org/3/library/re.html (visited on 06/01/2025).

[71] *DEAP documentation — DEAP 1.4.3 documentation*. URL: https://deap.readthedocs.io/en/master/ (visited on 06/01/2025).

[72] *Beautiful Soup Documentation — Beautiful Soup 4.4.0 documentation*. URL: https://beautiful-soup-4.readthedocs.io/en/latest/ (visited on 06/01/2025).

[73] *Hugging Face – The AI community building the future.* May 29, 2025. URL: https://huggingface.co/ (visited on 06/01/2025).

[74] *SentenceTransformers Documentation — Sentence Transformers documentation*. URL: https://sbert.net/ (visited on 06/01/2025).

[75] *Transformers*. URL: https://huggingface.co/docs/transformers/index (visited on 06/01/2025).