MTech - MCSE699J

INTERNSHIP II / DISSERTATION II REPORT ON

Detection of ethical principles based on argument using pre-Trained models

Submitted in partial fulfillment of the requirements for the degree of

Master of Technology

in

Computer Science and Engineering

(Artificial Intelligence and Machine Learning)

by

Tanul Khatri
23MAI0023

Under the Supervision of
Dr. Goutam Majumder

Assistant Professor Senior Grade 1

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

April 2025

DECLARATION

I, **TANUL KHATRI** hereby declare that the thesis entitled **"Detection of ethical principles based on argument using pre-trained models"** submitted to Vellore Institute of Technology (VIT), Vellore for the award of the degree of ***Master of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning)*** is a record of Bonafide work carried out by me under the supervision of **DR. GOUTAM MAJUMDER**, Assistant Professor Senior Grade 1, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore.

I further declare that the work reported in this project report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this instituteor any other institute or university.

Tanul

Signature of the Candidate

Place: Vellore

Date: 16/04/25

CERTIFICATE

This is to certify that the thesis entitled **"Detection of ethical principles based on argument using pre-trained models"** submitted by **TANUL KHATRI**, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore for the award of the degree of ***Master of Technology in Computer Science and Engineering (Artificial Intelligence and Machine Learning)***, is a record of Bonafide work carried out by him/her under my supervision during the period, **11th December 2024** to **16th April 2025** as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in partor in full, for the award of any other degree or diploma in this institute or any other Institute or University. The dissertation fulfills the requirements and regulations of the Institute and in my opinion, meets the necessary standards for submission.

Place: Vellore
Date: 15/04/2025

Name and Signature of the Guide~~Date:~~

Dr. Goutam Majumder

16.4.25
Internal Examiner

External Examiner 16/04/2025

Head of the Department
Program
MTECH CSE AI and ML
SCOPE

EXECUTIVE SUMMARY

The project, " Detection of ethical principles based on argument using pre-trained models," focuses on developing a system to automatically detect and categorize emotions in text, especially within arguments and opinion-based content. Recognizing that human values are often deeply intertwined with emotional expressions, the project seeks to analyze these nuanced emotions to uncover underlying values. Emotions play a pivotal role in shaping our beliefs, guiding decisions, and influencing how we interpret the world; this system thus aims to explore how specific emotional tones within text reveal motivations and value-driven perspectives. The project addresses the challenge of processing and accurately identifying complex emotional cues, particularly in argumentative discourse, where emotions like empathy, anger, or curiosity might subtly influence value judgments.

Leveraging machine learning and deep learning approaches, including transformer-based models, this project processes text data through a series of analytical steps, such as data preprocessing, feature extraction, and model training, to achieve precise emotion classification. Each step has been tailored to capture the depth of emotional content, allowing the model to detect both primary and subtle emotional layers within arguments. This granular analysis aids in understanding the complex emotional and psychological patterns behind various value systems, offering a richer understanding of human behavior and cognition.

The project has broad implications, potentially benefiting fields like social research, behavioral sciences, and human-computer interaction by promoting a more nuanced view of human values in digital communication. The system offers a foundation for applications that aim to foster empathy and awareness in interactions, making it possible to detect and reflect on the emotions and values underlying everyday discourse. This work aspires to support a more informed, value-conscious digital landscape, encouraging individuals and systems alike to appreciate better the emotional dynamics in human arguments and conversations.

The project, " Detection of ethical principles based on argument using pre-trained models," focuses on developing an automated system to classify and analyze emotions within argumentative text to reveal underlying human values. Emotions play a vital role in shaping individual beliefs and actions, often guiding the way people interpret events, express opinions, and make decisions. This system employs machine learning and deep learning methodologies, particularly transformer models, to analyze textual data and detect nuanced emotions in arguments. By capturing subtle emotional tones like empathy, anger, and curiosity, the model can offer insights into the values and motivations driving individuals' perspectives.

Through a combination of data preprocessing, feature extraction, and model training, the

project aims to achieve high accuracy in classifying emotions, providing a framework for understanding complex emotional patterns within argumentation. This work contributes to fields such as psychology, sociology, and computational linguistics by offering a tool that promotes empathy and value-awareness in digital communication. The findings have practical implications for advancing human-computer interaction, fostering informed discourse, and promoting value-conscious AI applications, setting the stage for more responsible and emotionally aware systems.

Keywords: *Classification, Logistic Regression, DNN, LSTM, Emotion Detection, Text Classification, Human Values, Natural Language Processing.*

ACKNOWLEDGEMENTS

With immense pleasure and a deep sense of gratitude, I wish to express my sincere thanks to my supervisor ***Dr. Goutam Majumder,*** Assistant Professor Senior Grade 1,School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore without his motivation and continuous encouragement, this research would not have been completed.

I am grateful to the Chancellor of VIT, **Dr. G. Viswanathan**, the Vice Presidents, andthe Vice Chancellor for motivating me to carry out research at the Vellore Institute of Technology.

It would be no exaggeration to say that the Dean of SCOPE, **Dr. JAISANKAR N**, was always available to clarify any queries and clear the doubts I had during the course of my project.

I would also like to acknowledge the role of the HoD, **Prof. J N Swathi**, who was instrumental in keeping me updated with all necessary formalities and posting all the required formats and document templates through the mail, which I was glad to have had.

Finally, I would like to thank **Vellore Institute of Technology** for providing me with infrastructural facilities, a flexible choice, and for supporting my research and execution related to the dissertation work.

Place: Vellore **NAME OF THE STUDENT**

Date: **TANUL KHATRI**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ML	-	Machine Learning
AI	-	Artificial Intelligence
EDA	-	Exploratory Data Analysis
NLP	-	Natural Language Processing
LR	-	Logistic Regression
OVR	-	One-vs-Rest
BoW	-	Bag-of-Words

Chapter 1

INTRODUCTION

1.1 Overview

Human values have long been studied in the social sciences by (Schwartz et al., 1994) and formal argumentation frameworks (Bench-Capon., 2003), and are defined as enduring beliefs about what is desirable and worth striving for. Schwartz's Theory of Basic Human Values identifies ten core values, including security, self-direction, and benevolence, which are structured hierarchically and validated across cultures. These values often align or conflict with one another, which can result in interpersonal disagreement or even the formation of ideological and political divisions.

Alongside values, emotions shape how individuals evaluate behavior, interpret situations, and communicate opinions. Emotions like anger, joy, or guilt are not random but deeply tied to a person's context, needs, and personality traits. These emotional experiences are commonly expressed through language, and efforts to capture them have resulted in structured datasets such as GoEmotions, ISEAR, and DailyDialog, where human annotators have labelled thousands of sentences with specific emotion categories. Such taxonomies allow computers to detect patterns in language that reflect emotional states.

Understanding and modelling both emotions and values are crucial in the field of Natural Language Processing (NLP) and human-computer interaction (HCI). Machines that can interpret emotional cues are better equipped to engage empathetically with users, provide mental health support, detect toxic or threatening content, and respond to arguments in a way that aligns with the user's values. This requires not only detecting emotional language but also reasoning about the underlying human values embedded in argumentative text.

To fully understand this dynamic, one must also consider a person's moral grounding and the framing of arguments—as discussed by (Feldman et al., 2021) and (Alshomary et al., 2022) in the context of value-based reasoning, and by (Entman et al., 1993) and (Ajjour et al., 2019) in the context of how emotional and value cues are framed within language.

These perspectives suggest that the same argument can be interpreted in vastly different ways depending on the emotional framing, cultural background, or moral orientation of the audience.

This study builds on the motivation set forth by SemEval 2023 Task 4 (Kiesel et al., 2022), which introduces the challenge of automatically identifying emotionally grounded human values in argumentative text. The task reflects real-world complexity by using a multi-label setup, where each sentence can be associated with multiple values from a consolidated taxonomy of 54 core human values. These values are often expressed implicitly and vary across cultures, making the task both linguistically and socially challenging. The benchmark dataset released as part of the task includes over 5,000 manually annotated arguments across multiple regions, offering an invaluable resource for training and evaluating value-aware NLP systems.

For example, the sentence “People should have the freedom to express their beliefs without fear” reflects an emotion of gratitude or hope, and maps closely to the human value self-direction, which emphasizes independent thought and freedom of expression.

In this research, we focus on enhancing value classification by leveraging emotion detection as an intermediate layer of understanding. We explore how integrating predicted emotional states into the input can help transformer-based models better predict human values, especially when those values are subtly implied. By combining insights from emotion modelling, value theory, and deep learning, this work aims to move one step closer to making AI systems more ethically informed, emotionally aware, and human-aligned.

SemEval’s value labels used in this study represent a broad range of human priorities, including:

- **Self-direction (thought, action)** – independent thinking, freedom of choice.
- **Stimulation** – seeking excitement, novelty.
- **Hedonism** – pursuing pleasure or enjoyment.
- **Achievement** – striving for success and competence.
- **Power (dominance, resources)** – influence or control over others and material possessions.

- **Face** – maintaining social status and respect.
- **Security (personal, societal) –** safety and stability.
- **Tradition** – respect for customs and cultural or religious heritage.
- **Conformity (rules, interpersonal)** – adherence to social norms or avoiding harm to others.
- **Humility** – modesty and downplaying personal importance.
- **Benevolence (caring, dependability)** – concern for others' welfare and being reliable.
- **Universalism (concern, nature, tolerance)** – understanding, protecting the environment, and accepting others' views.

1.2 Objectives

The objective of this project is outlined below:

- To develop an advanced emotion detection system that accurately identifies and analyses the emotions embedded in text-based inputs.
- To gain deeper insights into the segments and motivations that influence human values by understanding emotional nuances and patterns in natural language.

Emotion detection can enhance the human-computer interaction, it enables systems to respond in more emotional and human manner. It plays important role in customer service enhancement, content recommendations, and user experience. It also contributes to mental health monitoring, sentiment analysis, social media trend analysis, and creation of educational tools.

By recognizing emotions in text, businesses, healthcare platforms, educational applications, and social platforms can deliver more empathetic, intelligent, and human-centred experiences. Emotion detection also supports ethical AI development by allowing systems to better understand human perspectives, which is particularly important in applications such as argument mining, conflict resolution, and value-based reasoning.

1.3 Organization of the Report

A thorough grasp of the research on automated news classification for identifying and thwarting manipulative content is provided by the structure of this report. From the fundamental ideas to the final analysis and conclusions, every section addresses important facets of the research. This is the arrangement of the chapters.

Chapter 2: Literature Review: In this section pertinent research on automated news classification misinformation detection and manipulation techniques is reviewed. Along with sub-sections that highlight important insights and a summary to compile findings it also includes discussions of related works and earlier research in the field.

Chapter 3: Proposed Method – This chapter outlines the system's suggested methodology. It starts with a synopsis of the dataset that was used and a brief introduction. Along with discussing the shortcomings of current solutions and providing specifics on the design and module descriptions of the suggested system the chapter also introduces the architecture of the system. Finally, it describes the methods used in the classification procedure.

Chapter 4: Experimental Setup – This includes the technical configuration instruments and procedures used to put the system into use and assess its performance.

Chapter 5: Result Analysis and Discussion – In-depth analysis and discussion of the system's performance accuracy and efficacy in classifying manipulative news content are provided in this chapter which also presents the experiment findings.

Chapter 6: Conclusion and Future Improvements: This last chapter includes a summary of the study's main conclusions a discussion of its contributions and recommendations for future developments and possible lines of inquiry.

Chapter 2

LITERATURE REVIEW

Kiesel et al. [1] [2] has employed taxonomy of 20 values and their associated 54 values called level 2 and 1, for identification of human values behind the argument which were collected from 6 different source religious texts, political discussions, free-text arguments, newspaper editorials, and online democracy platforms. Previously authors contributed a dataset of 5270 arguments from 4 different geographical cultures: Africa, China, India, and USA. Bert, SVM and baseline were used for the identification, among which Bert Performed the best.

Papadopoulos et al. [3] had 3 different approaches in ANN, single multi-label classifier with sigmoid F1 Loss, Single Multi-label classifier that utilizes Bert base uncased and Bert's Large uncased English model, and Siamese Network, all three employees Bert and Roberta models. Ma et al. [4] has employed Roberta Large Model, ReWriteArgumentDataset, CB-NTR Loss function, and Ensemble Model. Hemati et al. [5] proposed a Label Graph Architecture which uses DeBerta and Adversarial Training. Monazzah et al. [6] designed a multi-label classifier to determine the value of the argument by using 2 multi-label classifiers based on Roberta and Bert. Jafari et al. [7] has used dataset which contains 9324 arguments and employed GoEmotion model on the which is already trained on multiple emotion dataset and it fine-grades the emotions into 3 groups +, - and neutral, which suggests the potential valuable features for enhancing the performance of human value detection algorithm. GoEmotion was able to surpass Baseline and Bert. Saha et al. [8]have used an ensemble of 3 models which are an entailment-based model for determining human values based on their descriptions, Roberta-based classifier that predicts the set of human values based from an argument, and a Roberta-based classifier to predict a reduced set of human values from an argument. Two-phase approach is used for training in which Roberta-base is pretrained on MultiNLI model and then optimized using AdamW. Zaikis et al. [9] have used a Transformer-based language model which utilizes second-phase pre-training in a one-versus-all (OVA) setting which allows the underlying LM to better represent the task in the embedding space and OVA improves the performance.

Machova et al. [10] has used multiple approaches on their Kaggle dataset which contains 20000 posts from different conversational content of social networks. Data pre-processed using tokenizer Api and padding. Different approaches were involved: Lexicon approach, Naïve Bayes, SVM, and Conv1D + LSTM. Deng et al. [11] has used ECG, touch sensors, and accelerometers to model emotions, RNN-LSTM, Bi-LSTM, ALSTM, ABLSTM, DBLSTM on the custom data by collecting data taken by random people using a band to collect 3-D accelerometer data. Jafari et al. [12] has performed the analysis on the semeval dataset.

Bandhakavi et al. [13] to identify emotions have used Lexicon model on multiple datasets, different combinations of features are considered, Total Emotion Count, Total Emotion Intensity, Max Emotion Intensity, Graded Emotion Count, Graded Emotion Intensity, and Contextual Features.

Chenna et al. [14] analyzed the twitter dataset by collecting random tweets on a keyword, collecting user information, collecting comments/replies on each tweet, collecting user information of the commenters, Preprocessing was performed.

Previous research in emotion detection has focused on building models to classify human values based on argumentative texts, particularly from SemEval 2023 Task 4. These studies grouped values into 20 broad categories and 54 specific labels, derived from diverse sources such as religious texts, political debates, and social media. Various approaches were explored, including the use of transformer-based models like BERT and RoBERTa, enhanced through techniques such as data preprocessing, data augmentation, and ensemble learning. The Label Graph Transformer (LG-Transformer) was employed to improve classification performance by capturing inter-value relationships, while other methods incorporated emotion detection to enrich the interpretation of value expressions. Grounded in Schwartz's Theory of Basic Human Values, these studies emphasize the cross-cultural nature of values and suggest future directions such as Chain-of-Thought (CoT) prompting for deeper contextual understanding.

2.1 Limitation of Existing System

The existing systems struggle to accurately detect emotions, which in turn makes it difficult for machines to truly understand human emotional expressions. These models are highly dependent on high-quality data, which is often difficult to obtain and maintain. Emotions like anger are particularly challenging to interpret compared to others, further complicating the detection process.

Another limitation is the system's inability to identify a complex range of emotions, such as blended or mixed feelings. Additionally, the ambiguity of natural language hinders the model's ability to interpret meanings behind words, idioms, and cultural references, leading to inaccuracies. Humor, sarcasm, and contextually complex language are especially difficult for machines to comprehend.

From a practical standpoint, the cost of building and training these models is high, as they require powerful GPUs, which are often expensive. Moreover, these systems face challenges in providing multilingual support, struggling with the nuances of diverse and complex languages.

There are also privacy concerns, particularly in sensitive areas like mental health monitoring or personal feedback analysis. Finally, the maintenance costs of such NLP systems are significant, as they require regular updates and retraining to stay effective and relevant.

Chapter 3

PROPOSED APPROACH

3.1 Brief Introduction

The proposed approach for emotion detection consists of comprehensive approach of data augmentation, data merging, model building, and iterative refinement. Two datasets are used, first is the semeval dataset, which consists of sentences and value labels and the second is a merged dataset compiled from 3 well-known datasets: GoEmotion [15], ISEAR [16], and DailyDialog [17]. The combined dataset was balanced by Upsampling and downsampling to maintain a consistent number of sentences per emotion value in the dataset and let the model train without any problems. Bert-based emotion classifier was trained on the merged dataset to perform multi-label classification across 32 emotion categories. The trained emotion model was then applied to semeval dataset with had no emotion annotations. To ensure every sentence in the semeval dataset had one emotion tag, top-k fallback strategy was used. To establish a baseline, Logistic regression was first used on semeval dataset. After setting up baseline, to further increase the performance, Bert-embeddings were extracted using Bert-base-uncased on semeval dataset, which were used with Logistic Regression which increased the performance of the model. For further enhancement, a DNN model was also explored using BERT embeddings, which led to a slight improvement in performance. Using the merged dataset and semeval datasets, two classification models were trained: model1 was trained on using only the sentences from the semeval dataset and model 2 was trained on both sentences and emotions combined. Both value classification models were built using DistilBERT (distilbert-base-uncased), fine-tuned for the multi-label task. Model 2 showed further improvement which confirming that emotional context improves value label prediction.

3.2 Dataset Description

The datasets used consists of textual data and emotion labels, which form the foundation for the training and testing of the model for prediction. Two datasets are used, first is obtained from the semeval task 4, which includes sentences compiled from multiple sources such as blogs, literatures, religious texts, text entries, etc. Each sentence is labelled with a set of human value labels. Semeval dataset contains 38 value labels: self-direction attained, stimulation constrained, Hedonism attained, etc. The second dataset is created by merging three well-known datasets: GoEmotion, ISEAR, and DailyDialog. The datasets are merged and pre-processed, followed by balancing them by Upsampling and downsampling to make sure every emotion label has sufficient number of sentences. The merged dataset contains 32 emotion values like: Joy, Anger, Sadness, etc.

Go-emotions: GoEmotions dataset is manual dataset collected by Google. GoEmotion dataset consists of 58,000 English Reddit comments labeled with 27 fine-grained emotion categories and a 'neutral' class. This dataset contains both basic and complex emotions like Joy, Sadness, Admiration, etc. Each comment is associated with atleast one emotion and even more, which makes it suitable for multi-label classification.

ISEAR: Internation Survey on Emotion Antecedents and Reactions (ISEAR), is a dataset created by collecting responses from people across 26 countries. People from these countries were asked to describe the emotions they had in a certain situation which provided with seven fundamental emotions: Joy, Sadness, Anger, Disgust, Shame, and Guilt. This dataset contains over 7,500 texts.

DailyDialog: DailyDialog is a high-quality, multi-turn dialog dataset which contains around 13,000 conversations which reflect daily communications topics. Each dialog is annotated with emotions labels like: happiness, anger, surprise, sadness, no emotion, etc. The sentences topics varies from daily-to-daily life scenarios. This dataset is widely used for emotion detection for its natural flow and clear emotion annotations.

Text: The semeval dataset contains above 100,000 sentences after balancing them. Each entry is identified by a combination of Text-ID and Sentence-ID. The Text column contains actuals sentences. There are 1,604 unique Text-IDs and 83 unique Sentence-IDs. The merged emotion dataset contains 74,897 sentences and 32 unique emotion categories. It also includes three different identifier columns: Text-ID, Sentence-ID, and Sentence. Text-ID and sentence-ID contains 74,897 and 43,410 unique ids respectively.

Labels: The semeval dataset provides detailed annotations for each sentence, associating with specific emotional labels. Each text is identified by a unique text-ID and Sentence-ID. There are total of 38 emotion labels (19*2(attained and constrained)). Labels use values of 1, 0, 0.5 to indicate presence or absence of emotions, where 1 signifies the sentence contain that emotion, 0 signifies sentence does not contain that emotion, and 0.5 signifies the sentence neither fully present nor absent. The merged emotion dataset provides detailed annotations for each sentence, with 32 unique emotions.

Before training, several Data augmentation and pre-processing techniques are applied on both the datasets. Section 3.4.4 mentions different data pre-processing techniques. Bert embeddings were generated for each sentence using the Bert-Base-uncased model to capture textual meaning. The datasets were split into training, validation, testing with ratio of 80: 10: 10. Due to natural Distribution differences of emotional labels in the datasets varies, some emotions are more frequent than the others which introduces the class imbalance. This can affect the model's ability to not generate and may bias it toward majority classes, a challenge addressed through dataset balancing and model architecture tuning.

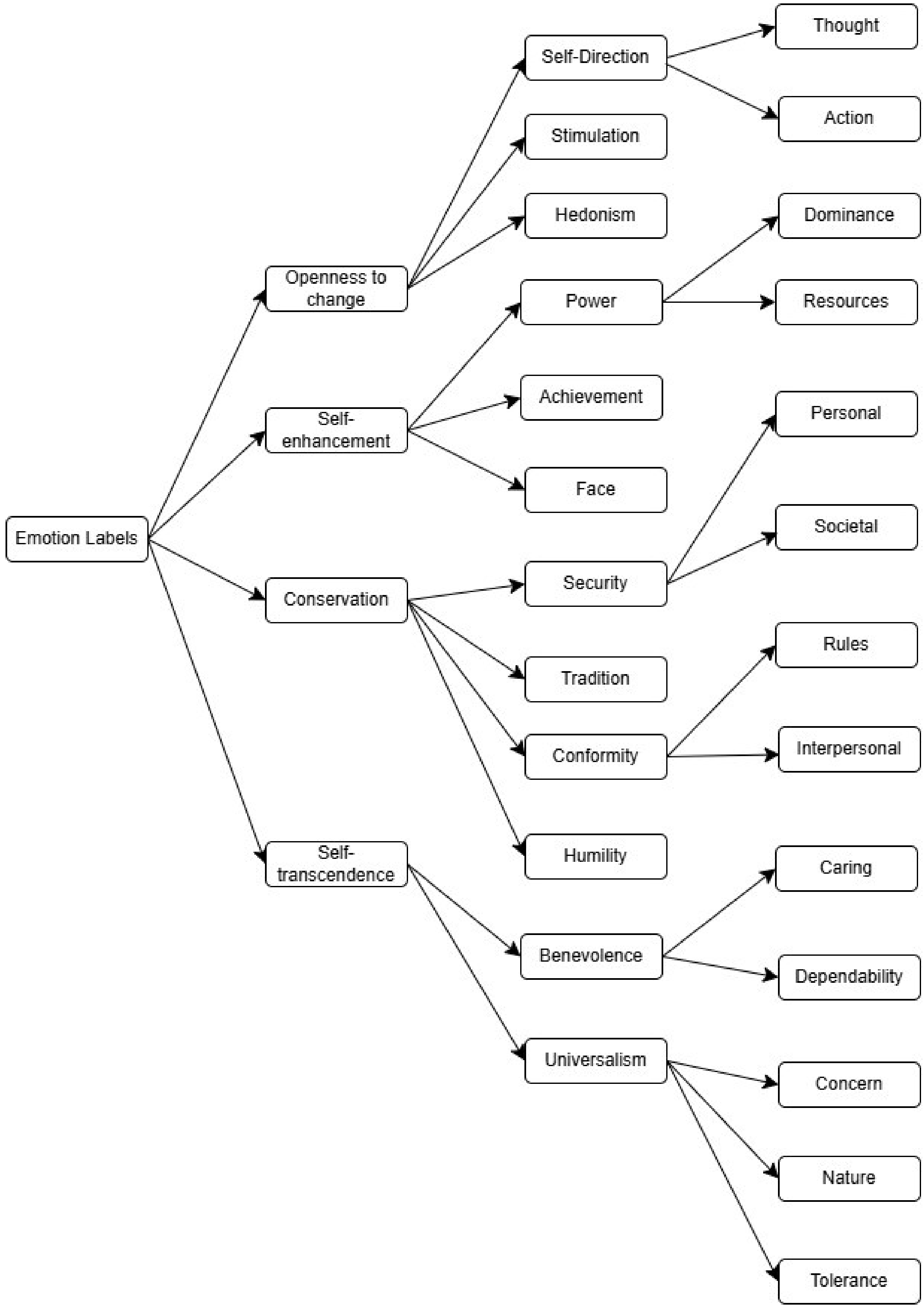

Figure 3.1 Value Labels dataset

Figure 3.1 illustrates a tree structure that shows the emotion labels dataset into four main categories: Openness to change, Self-enhancement, Conservation, and Self-transcendence, further divided into sub-branches like Self-direction under Openness, which also splits into different emotions. The tree structure more efficiently shows how the emotions are divided into parts and then sub-parts simplifying the understanding of the label dataset

Table 3.1 Sentence Description

Text-ID	Sentences-ID	Text
EN_001	1	Hispanic Voters Are Losing Faith In the Democracy.
EN_001	2	The support of Hispanic voters at the midterms.
EN_001	3	U.S. President Joe Biden speaks to employees.
EN_001	4	Legal, Not Illegal, Immigration.
EN_001	5	This marks the lowest approval rating of any demographic group.

Table 3.1 presents a sample of the dataset used. The table includes columns for the Text-ID, Sentence-ID, and the Text. Each row represents unique sentences, and combination of Text-ID and Sentence-ID. The Text-ID specifies the source document, while the Sentence-ID indicates the sentence's sequence. The Text column contains the actual content of the sentences. This structured format provides a foundation for conducting sentiment and emotion analysis across various texts.

Our approach comprises of machine learning and deep learning techniques. It begins with logistic regression on raw text to establish a baseline, followed by fine-tuning a more advanced transformer-based model to improve detection accuracy. Different approaches used are Deep Neural Network (DNN) and Value classifier Model. This methodology supports the broader goal of using artificial emotional analysis to explore emotional patterns and human behavior, ultimately deepening our understanding of how emotions shape values across various contexts.

Table 3.2 Labels Description

Text-ID	EN-001	EN-001	EN-002	EN-002	EN-003
Sentence-ID	1	2	3	4	5
Security: societal	0	1	0.5	1	0
Security: personal	0	0	0	0.5	0
Power: resources	1	1	0	0	1
Power: dominance	0.5	0.5	1	0	0.5
Achievement	0	0	0.5	1	0.5
Universalism: concern	1	1	0	0	1
Universalism: nature	0.5	0.5	1	0	0.5
Universalism: tolerance	1	1	0.5	0	0
Benevolence: caring	0	0	0	0.5	1
Benevolence: dependability	1	1	1	1	1
Conformity: rules	0.5	0.5	1	0.5	0
Conformity: interpersonal	1	1	0	1	1
Self-direction: thought	0.5	0	0	0	0.5
Self-direction: action	1	1	1	0	1
Stimulation	0	0.5	0	1	0
Hedonism	0	0	1	0	0
Tradition	1	1	0	1	1
Humility	0	0	0.5	0	0.5
Face	0	1	0	0.5	1

Table 3.2 provides an overview of the emotional attribute annotations for a subset of sentences from the dataset, identified by Text-ID and Sentence-ID. Each row represents a specific emotional attribute, and each column corresponds to a particular sentence, with the associated Text-ID (e.g., "BG-001", "BG-002", "BG-003") and Sentence-ID (e.g., 1, 2, 3, 4, 5) indicating the sentence's unique identifiers. The table captures the presence and intensity of various emotional attributes, represented by values of 0, 0.5, or 1. A value of 1 indicates the full presence of an emotional attribute, 0 denotes the absence, and 0.5 suggests partial association. This structured representation allows for detailed analysis of how different emotions are distributed and expressed across the text samples.

Table 3.3 Sentence Description of Merged Dataset

Text-ID	Sentence-ID	Text
Train_001	1	I can't believe how rude he was at the party
Train_001	1	The movie made me feel incredibly happy and warm inside
Train_001	2	I don't know what to say anymore, I'm just confused
Train_001	1	It was such a joyful experience to be there
Train_001	2	I feel disappointed with how things turned out
Train_001	1	He was shocked when he heard the news
Train_001	3	Her performance left everyone in awe and admiration

Table 3.3 provides sample entries from the merged emotion dataset, which combines sentences sourced from multiple well-known datasets. Each entry is uniquely identified by a Text-ID and Sentence-ID, ensuring traceability across training (train_), validation (val_), and testing (test_) splits. The Text column contains emotionally rich natural language statements used for multi-label emotion classification.

Table 3.4 Number of sentences per language

Language	**Number of Sentences**
BG	4000
DE	5500
EL	4500
EN	4800
FR	3000
HE	5000
IT	3500
NL	6200
TR	6200

Table 3.4 describes the number of sentences for each language in the dataset, revealing significant differences in representation. NL (Dutch) and TR (Turkish) have the highest counts, each with 6200 sentences, suggesting strong coverage in these languages. In contrast, FR (French) is underrepresented with only 3000 sentences, which may impact model performance for French text. Other languages, such as DE (German) with 5500 and HE (Hebrew) with 5000 sentences, have moderate representation. BG (Bulgarian), EL (Greek), EN (English), and IT (Italian) range between 3500 and 4800 sentences. The language codes represent: BG (Bulgarian), DE (German), EL (Greek), EN (English), FR (French), HE (Hebrew), IT (Italian), NL (Dutch), and TR (Turkish). The uneven distribution highlights potential challenges in model training, requiring strategies like data augmentation or weighted loss functions to ensure balanced performance across languages.

Table 3.5 Merged Dataset Distribution by Source

Dataset Source	**Number of Sentences**
GoEmotion	58700
DailyDialog	10800
ISEAR	5400
Total	**74900**

Table 3.5 summarizes the source-wise distribution of sentences in the merged dataset. The dataset comprises a total of 74,900 sentences, with the majority sourced from GoEmotion (58,700), followed by DailyDialog (10,800) and ISEAR (5,400). This diverse mix ensures that the model is trained on both formal and informal text types—ranging from conversational dialogues (DailyDialog), emotionally rich statements (ISEAR), to crowd-sourced annotations (GoEmotion). This improves the robustness and generalizability of the emotion detection system across various linguistic and contextual scenarios.

Table 3.6 Merged Emotion Value Dataset

Admiration	Amusement	Anger	Annoyance
Approval	Caring	Confusion	Curiosity
Desire	Disappointment	Disapproval	Disgust
Embarrassment	Excitement	Fear	Gratitude
Grief	Guilt	Happiness	Joy
Love	Nervousness	Neutral	No Emotion
Optimism	Pride	Realization	Relief
Remorse	Sadness	Shame	Surprise

Table 3.**6** illustrates the structure of the Merged Emotion Value Dataset, which lists the emotions present in the combined dataset. This dataset was created by merging three well-known sources: GoEmotion, ISEAR, and DailyDialog. It includes 32 distinct emotions that reflect the most common and frequently observed emotional states expressed by individuals in day-to-day life.

Table 3.7 Top 10 Emotions

Emotion Label	**Total Score**
Security: societal constrained	2100
Achievement attained	1900
Security: societal attained	1800
Conformity: rules attained	1700
Universalism: concern attained	1600
Power: resources attained	1500
Power: dominance attained	1450
Power: dominance attained	1400
Conformity: rules constrained	1300
Stimulation attained	1200

Table 3.7 above provides a summary of the top 10 emotional attributes identified in the dataset, ranked by their total score, which represents the frequency or intensity of each emotion expressed in the text. Security: societal constraints have the highest score of 2100, indicating that societal security concerns and limitations are the most prominent themes. Achievement attained and Security: societal attained followed by

scores of 1900 and 1800, reflecting a focus on personal and societal accomplishments. Conformity: rules attained and Universalism: concern attained highlight the importance of social norms and concern for others, with scores of 1700 and 1600. Themes of authority and control are evident in Power: resources attained and Power: dominance attained, scoring 1500 and 1450, respectively. Lower-ranked attributes like Self-direction: action attained, Conformity: rules constrained, and Stimulation attained still contribute significantly to the emotional landscape but to a lesser degree.

Table 3.8 Top 10 Most Frequent Emotions in Merged Dataset

Emotion Label	Total Count
Joy	7400
Neutral	6950
Anger	6450
Sadness	6200
Disgust	6100
Fear	5900
Surprise	5800
Embarrassment	5600
Amusement	5450
Gratitude	5300

Table 3.8 highlights the top 10 most frequent emotions in the Merged Emotion Dataset. This table provides insight into the emotional distribution across the combined dataset sourced from GoEmotion, DailyDialog, and ISEAR. The most dominant emotion observed is Joy with 7,400 instances, followed by Neutral and Anger. The presence of high-frequency emotions like Sadness, Disgust, and Fear reflects a diverse range of emotional expressions captured in the dataset. This distribution ensures the model is exposed to a wide spectrum of emotions, enabling it to generalize well during classification.

The text sources for this project span multiple domains and datasets, including religious texts, political discussions, free-text arguments, newspaper editorials, and online democracy platforms from the SemEval dataset, as well as emotional expressions derived from the merged GoEmotions, DailyDialog, and ISEAR datasets. Emotions are integral to shaping human values, as they influence how individuals interpret events, interact with others, and make decisions. Values, which serve as core beliefs guiding behavior and perceptions, are often shaped by a combination of logical reasoning and emotional experiences. This project leverages fine-grained emotion analysis to extract subtle sentiments from text, enhancing the understanding of underlying value systems and distinguishing complex emotional states. For example, differentiating between anger stemming from perceived injustice and excitement over new perspectives can yield meaningful insights into a person's values and motivations.

3.3 Architecture of the Existing Method

3.3.1 Proposed System

The proposed system aims to identify emotions in the text by using NLP and deep learning techniques, focusing on transformer-based models to capture the emotional content of text. This approach requires the emotional analysis, acknowledging that emotions significantly influence human values, perceptions, and interactions. This system processes data to identify the emotion expressed, providing deeper insights into the underlying emotions and thought processes.

Input Data and Pre-processing: Each sentence is pre-processed by taking the labelled dataset. During pre-processing, these texts are analyzed to identify the emotion(s) in each sentence, making the task to a multi-label classification task easier. Each sentence is then passed through text pre-processing steps, like removing punctuation, lowercasing, and eliminating unnecessary tokens.

Baseline Model: To establish a baseline performance Logistic regression is used on pre-processed data. This model provides a baseline for comparison, by showcasing how well simpler linear models capture emotion versus more complex models. The baseline model helps gauge the starting point for accuracy and informs the effectiveness of more complex methods. The Logistic Regression model was trained using Log-Loss as the objective function and run for 1000 iterations to ensure convergence. Support Vector Machine (SVM) and Naïve Bayes models were also introduced; however, due to their poor performance, they were not adopted for further evaluation. Instead, Logistic Regression was chosen to establish the baseline, given its better performance and interpretability.

SVM: Support vector Machine (SVM) was also used using pre-processed data and Bert Embeddings with addition of NRC (National Research Council). NRC are the common features in emotion detection which are describe the emotions to the machine. SVM did not perform well enough against Logistic Regression. Due to its lesser performance, SVM was not included in final results.

Naïve Bayes: In addition to SVM, Naïve Bayes was also employed on same pre-processed data and Bert Embeddings with addition of NRC. Naïve Bayes did not perform well enough against Logistic Regression, so it was not included in final results

Transformer-Based Model: The primary model within the proposed system is a transformer-based fine-tuned BERT model. Transformers models are effective at identifying context-dependent emotions in text. BERT captures both word-level and sentence-level contextual information, making it a good choice to handle the emotional expression that might not be readily apparent. Fine-tuning BERT on the target dataset enhances its ability to detect a wide range of emotions, thereby aligning it more closely with the specific goals of our classification task. The model was trained using Log-Loss as the loss function, with L2 regularization to prevent overfitting, and ran for 1000 iterations to ensure sufficient convergence.

DNN: A Deep Neural Network (DNN) is explored in the proposed system, DNN is applied over Bert embeddings. From Bert-base-uncased model contextual sentence representations are obtained then the embeddings were flattened to create a fixed-length vector input. The model was trained for 10 epochs with a batch size of 32, using BCEWithLogitsLoss as the loss function and the Adam optimizer. The input vectors were passed through a series of fully connected dense layers, with ReLU activation functions applied to introduce non-linearity. To reduce the risk of overfitting, dropout layers were incorporated between the dense layers. The final dense layer uses a sigmoid activation function to allow muti-label emotion classification. This architecture aimed to capture high-level abstract features from the Bert-embeddings while maintaining computational efficiency.

DistilBERT VC: To get further improve the performance a merged dataset was used, a DistilBERT-base-uncased model was fine-tuned on the merged emotion dataset which handled 32 emotion classes using Sigmoid activation for multi-label classification. To tag the sentences from the semeval dataset emotion classifier was used as the semeval dataset did not have any emotion annotations. To make sure every semeval sentence received atleast one predicted emotion label, a top-k fall back method was used alongside per-class thresholding. The emotion classifier was trained for 5 epochs with a batch size of 8, using BCEWithLogitsLoss as the loss function. Following this, a value classifier was developed using semeval dataset. The first model was developed by training the model on the semeval dataset to predict the value labels using only sentences and leaving the emotions for next model. The second model was used on the semeval dataset just instead of only using texts, emotions were also used to enhance the performance and to test if adding emotions improves value prediction, this allowed the model to use both text and emotion labels. Both models were trained using the same configuration: 5 epochs, batch size of 8, BCEWithLogitsLoss, and the AdamW optimizer. After comparing both the models it was determined that adding emotions with text improved the performance.

Output and Model Interpretability: For the final output, emotion label is predicted for each sentence, representing the emotion detected in the text. Also, probability scores are provided by the system for possible emotion, offering a measure of confidence for each classification.

Advantages and Potential Impact: The proposed system works as valuable tool by capturing complex emotions. The Value classifier Model approach enables the model to capture emotional cues, providing reliable emotion detection that could aid in understanding human values, decision-making, and behavior in data-driven research. This system thus contributes to a more understanding of emotional expression, making a for deeper insights into the emotion detection of human interaction and values.

3.3.2 Description of the Module

- Data preprocessing: Removes all the whitespaces and lowercasing is performed on data. Data splitting: Separating data into different sets of training, validation, and testing.
- Feature Extraction: Uses techniques like Bag of Words (BoW) to transform text into numerical data and feature extraction.
- Model training: Uses training data to train model.
- Assessment: F1 score metric is used for testing model's performance.

3.4 Methodologies Adopted

3.4.1 Project Overflow

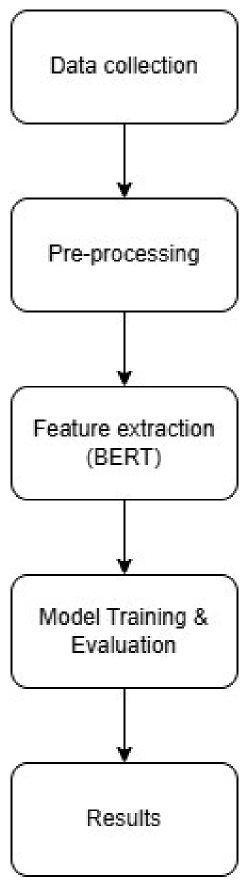

Figure 3.2 Workflow of proposed system

Figure 3.2 illustrates a workflow diagram of the proposed system. The steps include data collection, where the dataset is first gathered and passed to the preprocessing stage. Here, various data cleaning techniques are applied to make it suitable for model training. After preprocessing, features are extracted using the BERT-base-uncased model in the form of vector embeddings. These BERT embeddings are then used to train and evaluate the model, resulting in performance metrics such as F1-score, accuracy, and other relevant evaluation measures.

3.4.2 Architecture Diagram:

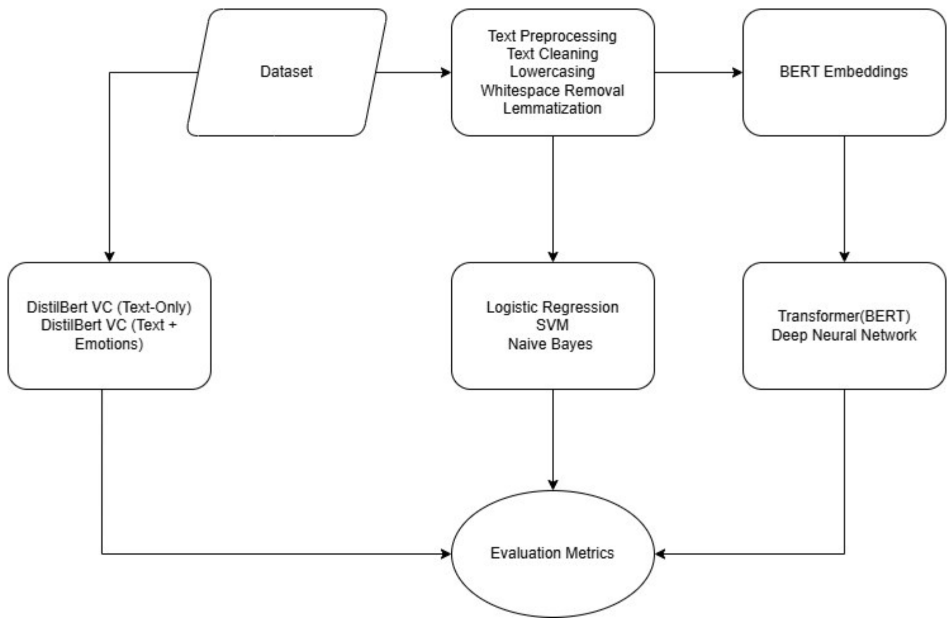

Figure 3.3 Architecture of Proposed System

The architecture shown in Figure 3.3 is detailed above, outlining the workflow of the emotion detection system. In step one, text data is taken from the semeval 2023 task 4 and a dataset is created using three well-known datasets: Go-Emotion, ISEAR, and DailyDialog which are then merged to create a single dataset. Different Data pre-processing techniques are performed on the data like text cleaning, lowercasing, whitespace removal, and lemmatization, to make it better for analysis.

Next, classification is done by models which are trained on pre-processed data, beginning with a simple Logistic Regression setting a baseline, followed by Bert Embeddings are extracted from pre-processed data and used in logistic regression for better accuracy and score. For better results DNN was used which provided better accuracy and score. Subsequently, DistilBERT VC (text only) and DistilBERT VC (text + Emotions) were developed - model 1 and model 2 were developed which gave better results especially model 2. Finally, the system's is evaluated using the F1-score, balancing precision and recall to ensure accurate and reliable emotion detection. This approach combines traditional and deep learning methods for comprehensive emotion analysis.

3.4.3 Exploratory Data Analysis:

Exploratory data analysis (EDA) is an important initial step in the data analysis process and using machine learning tools to help accomplish tasks. It involves mathematical and graphical approaches aimed at understanding the inherent organization, design, and deviations within the data. However, before modeling is performed EDA is beneficial because it allows data scientists to formulate questions and proposes relationships looks for unusual values and check for data integrity issues including but not limited to incomplete and erroneous data. The insights from this process enhance predictive modeling by guiding feature selection and engineering as well as model selection.

Within the context of the Detection of ethical principles based on argument using pre-Trained models, EDA is a crucial step in preparing the dataset for emotion detection tasks. This process begins by investigating the properties of the dataset such as the number of sentences, number of emotion labels and values, distribution of emotion across dataset. By identifying the class distribution of genres, one could forecast the need for oversampling, under sampling, or class weights adjustment in the presence of class imbalance.

The following important stage involves the review of the actual text data. It involves divided all the given text into chunks for the purpose of collecting information and visualizing graphs in particular genres of literature. Word clouds and bar charts can be very helpful in understanding the most used words and if some are used more than others in certain genres. For example, how many sentences per language, and top labels present in the dataset, they would be quite useful for classification.

Text preprocessing procedures like stop words extraction, stemming and lemmatization should also be investigated in terms of their effect to the quality of the data. Moreover, examining the relationship between the sentences and its labels can be beneficial in uncovering valuable insights. For instance, it is possible that larger number of sentences may be more present for one or more labels or smaller number of sentences per label.

Additionally, text preprocessing techniques like removing stop words, stemming, and EDA also includes detecting any incomplete or contradictory information of the dataset which is usually associated with text datasets. As an illustration, the absence of some labels or null texts may present a barrier in training models and hence should be managed properly. In addition to this, checking for the duplicity of sentences can make sure that the model is trained on diverse datasets.

With regards to citing literature, it should be noted that EDA for Detection of ethical principles based on argument using pre-Trained models is not only the stage which prepares the data for modelling but also has a role in developing the target model by employing feature engineering and preprocessing as well as modelling itself. By conducting such procedures, it is possible to help oneself by enhancing the ability to evaluate and anticipate the dataset effectively during model development in order to achieve better end results. Word Clouds are generated for understanding the data better in terms of the secondary class.

3.4.4 Data Preprocessing:

Data preprocessing steps undertaken in this project were essential to ensure that the raw text data was cleaned, structured, and appropriately transformed to provide high-quality input for the emotion detection model. Below is a detailed explanation of the data preprocessing steps applied in the project:

Text Cleaning Removal of Special Characters and Numbers: Text data often contains unnecessary special characters, symbols, and numerical values. These were removed as they do not contribute meaningfully to emotion detection. For example, characters like @, #, or any numbers were stripped from the text.

Lowercasing: All text was converted to lowercase to standardize the data. This ensures that words like "Happy" and "happy" are treated as the same, reducing redundancy and inconsistency in the input data.

Whitespace Removal: Extra spaces between words were reduced to a single space, and leading or trailing spaces were removed to ensure clean input.

Tokenization: Texts are split into words and allotted a token, which helps in breaking down sentences into more easier units for analysis. This step is crucial for understanding the meaning and context of each word about the others.

Stop word Removal: Stop words, which are common words like "is," "the," "and," etc., were removed from the text. These texts don't contribute to the model training instead can create problems while training.

Lemmatization: Words were lemmatized WordNetLemmatizer [18] from the NLTK library to convert them to their base or root forms (e.g., "running" to "run" and "better" to "good"), which reduces the number of unique tokens, hence improving model's efficiency.

BERT Embedding Generation: BERT (Bidirectional Encoder Representations from Transformers) [19] was used to generate contextual word embeddings. Each sentence in the dataset was converted into a vector representation using pre-trained BERT models. These embeddings capture rich semantic meaning and context of the words and sentences, which is essential for emotion detection, as emotions are context-dependent. The BERT embeddings were used as features for training the model, allowing it to understand the subtle emotional cues within the text.

Label Encoding: The emotion labels were encoded numerically to help in the training of models. Each emotion label was mapped to a corresponding numerical value, which allows the model to perform classification based on these encoded labels.

Data Splitting: The dataset is divided into three parts, training (80%), validation (10%), and testing (10%), dividing the dataset into parts makes it helpful for model to train, validate and test the data for better accuracy and emotion detection.

Handling Missing Data: Any missing or incomplete entries are removed, which if not removed can cause problem in model training and affecting the model's performance.

Below is the example of text before and after pre-processing for better understanding

Original Sentence: “I can't believe how amazing this movie was! Totally blew my mind!!”

After Preprocessing: “believe amazing movie totally blow mind”

By performing these preprocessing steps, the dataset was transformed into a clean and structured form, ready for model training. These techniques ensure that the input data is of high quality, facilitating more accurate and efficient training of the emotion detection model. The preprocessing also reduces noise and irrelevant information, allowing the model to focus on the essential features necessary for emotion analysis.

3.4.5 Feature Engineering

feature engineering plays a vital role in transforming raw text data into meaningful representations suitable for models, by capturing the semantic and contextual information of text to enhance model’s performance and emotion detection.

BERT Embeddings: Contextual Word Representations: BERT Model is used to extract embeddings for sentences in the dataset. BERT is a state-of-the-art language model that captures the context of a word based on its surroundings, providing deep and nuanced representations.

Preprocessing for BERT: The data was tokenized and pre-processed to make it compatible with BERT model, which involved process like padding for providing a uniform length for efficient input into the model.

Embedding Extraction: BERT embeddings were generated for sentences, resulting in dense, high-dimensional vectors which encodes semantic meaning. These embeddings serve as the primary features for emotion detection, allowing the model to differentiate between emotions based on subtle linguistic cues.

Sentence Length and Word Count: Sentence Length: Number of characters in sentences are calculated to analyze the complexity and verbosity of the text. Sentence length can be an important feature, as longer sentences might convey more complex or mixed emotions.

Word Count: The number of words in each sentence was computed as an additional feature, which helps in understanding the distribution of sentence lengths and provide insights of emotional data.

Text Preprocessing: Lowercasing: All text was converted to lowercase to maintain uniformity and reduce the complexity of the vocabulary.

Stop word Removal: Even though BERT can handle stop words, for simpler baseline models, stop words are removed from the sentences.

Punctuation and Special Character Removal: Punctuations and special characters are removed from the sentences to make it more efficient and easier for model to detect emotions in sentences.

3.4.6 Classification Model

The objective of this project is to detect the emotions behind the given texts. To complete this task both machine learning and deep learning techniques like logistic regression and BERT embeddings, obtained from BERT-base-uncased model respectively are used. More techniques like DNN and value classifiers are used to enhance the performance.

BERT-base-uncased Embeddings: A pre-trained transformer model BERT-base-uncased is used to generate embeddings of each sentence. Bert embeddings capture the semantics of text and are particularly effective for emotion detection tasks due to their deep understanding of language context.

Feature Engineering: To enhance the embeddings for making them more informative for classification task, feature engineering techniques are applied.

Logistic Regression (LR) for Multi-label Classification: For setting the baseline, Logistic Regression is used for emotion detection. Provided emotion detection involves multiple overlapping emotions, we implement Logistic Regression in a multi-label classification setting.

$$y_i = \sigma(W_i^T X + b_i) \qquad (1)$$

Where:

X = input feature vector

$w_i b_i$ = weights and bias for i-th label

y_i = predicted probability for the i-th label

$\sigma(z) = \frac{1}{1+e^{-z}}$

Deep Neural Network: To improve upon the baseline, a Deep Neural Network (DNN) is used for emotion classification. The model consists of multiple fully connected layers with non-linear activation functions (ReLU), which allows it to capture complex patterns in the input text representations. The final layer uses sigmoid activation for multi-label classification, enabling the detection of multiple co-occurring emotions in a single sentence.

$$h_1 = f(w^{(1)}x + b^{(1)}) \quad (2)$$

$$h_2 = f(w^{(2)}h_1 + b^{(2)}) \quad (3)$$

$$Y = \sigma(w^{(3)}h_2 + b^{(3)}) \quad (4)$$

Where:

X = Input

$w^{(1)}, w^{(2)}, w^{(3)}$ = weight matrices

$b^{(1)}, b^{(1)}, b^{(1)}$ = bias vectors

$f(z) = \max(0, z)$ = ReLU activation

$\sigma(z)$ = sigmoid function

Y = predicted vector of probabilities for each label

Value Classifier: For predicting human values from textual arguments, a classifier is developed using the DistilBERT (distilbert-base-uncased) architecture. The model takes sentences as inputs and leverages the contextual representation of the [CLS] token, followed by a dense layer with sigmoid activation for multi-label classification. This setup enables the system to learn deep semantic relationships in text and accurately identify multiple human values (attained/constrained) present in argumentative discourse.

$$Y = \sigma(W \cdot H_{|CLS|} + b) \quad (5)$$

Where:

$H_{|CLS|}$ = contextual representation from the final transformer layer

W, b = weight and bias of the classification head

Y = multi-label prediction probabilities

σ = element-wise sigmoid activation

One-vs-Rest (OvR) Strategy: OvR approach is applied to handle the multiple labels, where LR classifier is trained for each emotion label. Each classifier independently predicts whether an emotion is present or absent in a given sentence.

Probability Thresholding: Logistic Regression outputs a probability score for each emotion, which is mapped to a binary classification based on a chosen threshold (e.g., 0.5). This means that an emotion is considered "present" if its probability exceeds the threshold.

Formula: The probability that a given instance belongs to an emotion class is computed as:

$$P(y = 1 \mid X) = \frac{nx}{1+e^{-(\omega.X+b)}} \quad (6)$$

Where:

P(y=1|X) is the probability of the outcome y=1, given the features X.

ω·X represents the dot product of the weight vector ω and the feature vector X.

b is the bias term.

e is the base of the natural logarithm.

During training, the weights (ω) are optimized to minimize log loss (cross-entropy loss), making Logistic Regression a strong choice for capturing linear relationships in the data.

Deep Learning with BERT-base-uncased: Fine-Tuning BERT: We explore fine-tuning the BERT-base-uncased model on our emotion detection dataset. By doing this, the model adapts to recognize patterns specific to our emotion labels. Fine-tuning

BERT enables us to capture deeper, context-rich representations and improve classification accuracy.

Multi-label Output Layer: A custom classification head is added to BERT, with multiple output nodes corresponding to each emotion label, configured for multi-label classification.

Objective Function: BERT's training is guided by a multi-label classification loss, where each label's prediction is treated independently. The loss function, typically Binary Cross-Entropy (BCE) loss, is minimized across all classes.

$$\text{BCE Loss} = f(x) = \frac{1}{N}\sum_{i=1}^{N}(y_i \log P_i + (1 - y_i)\log(1 - P_i)) \quad (7)$$

Where:

Y_i: True label for the i-th emotion.

P_i : Predicted probability for the i-th emotion.

N: Total number of labels.

Model Comparison and Evaluation Metrics: Macro F1 Score: Since the dataset includes multiple classes and imbalanced labels, the Macro F1 Score is used to evaluate the model's performance across all emotion labels, providing an average score without bias toward frequent classes.

Classification Report: Precision, recall, and F1 scores are calculated for each emotion to understand the model's ability to distinguish between different emotional contexts.

By comparing the performance of Logistic Regression with that of the fine-tuned BERT model, we aim to determine the most effective approach for emotion detection, balancing accuracy with computational efficiency.

3.4.7 Model Training

Model Instruction. We train the chosen model after the model selection process is finished. Using the scikit-learns train_test_split function the data is first divided into training and testing sets with 80% of the data going toward training, 10% toward testing, and 10% toward validation. After preparing our data we fit our model to the training set and assess performance using the test set.

3.4.8 Model Evaluation

Four key metrics – precision, Recall, F1-score, and Accuracy are used to evaluate the model's performance. These metrics enable us to assess the efficacy of the model by offering insights into different facets of classification performance. Precision quantifies the proportion of actual positive outcomes that match the predicted positive ones.

$$Precision = \frac{TP}{TP+F} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recal} \tag{10}$$

$$Accuracy = \frac{TP+FN}{TP+TN+FP+F} \tag{11}$$

Where:

TP – True Positive

TN – True Negative

FP – False Positive

FN – False Negative

When assessing model performance these metrics are crucial and higher scores typically correspond to more effective classification. Following evaluation, the trained model can be successfully applied to news genre predictions on unseen data. For this case, the F1 score is considered for most of the cases as it would be easy to compare the results with other state-of-the-art methods.

Chapter 4

EXPERIMENTAL SETUP

This project focuses on developing an emotion detection system that identifies and classifies emotions in text using an existing dataset. The dataset provides a comprehensive foundation for training and evaluating models, as it encompasses diverse text samples labeled with various emotional and psychological attributes.

Data Collection and Preparation

- The data preparation phase involved a series of preprocessing steps to ensure high-quality input for modeling:
- Preprocessing Steps: The text data was cleaned by removing stop-words, special characters, HTML tags, and email addresses to enhance model performance and standardize the input.
- Tokenization: The text was broken down into individual words or tokens, facilitating structured analysis and feature extraction.
- Data Split: The data was divided into training (80%), testing (10%), and validation (10%) sets using the train_test_split function from scikit-learn, enabling proper training and evaluation of models.

Libraries and Tools

- Several essential Python libraries and tools were used throughout the project:
- NLTK: Employed for text preprocessing tasks, such as tokenization and stop-word removal.
- Scikit-learn: Used for data splitting, feature extraction, and implementing traditional machine learning models.
- Text Vectorization: Technique like Bag-of-Words (BoW) is applied for simple numerical representations of text.
- BERT: For more advanced feature extraction, BERT embeddings is used to

capture the semantic meaning of words.

Feature Engineering

- To convert the text into a numerical format suitable for machine learning, three main feature extraction methods were utilized:
- Bag-of-Words (BoW): Represented the word frequencies in each text sample using a sparse matrix.
- BERT Embeddings: These techniques provided dense word representations, improving the model's ability to capture semantic relationships and emotional nuances.

Model Selection

- Various machine learning models were explored to identify the most effective approach for emotion classification:
- Baseline Models: Logistic Regression was implemented as baseline models for its simplicity and interpretability. Support Vector Machin (SVM) and Naïve Bayes were also experimented with, they were not selected as baseline model due to their inability to deliver satisfactory performance on the task.
- Advanced Models: BERT-base-uncased model was used to improve accuracy and capture contextual dependencies in the text.
- Value Classifier: Transformer models like DistilBERT (distilbert-base-uncased) were fine-tuned as value classifiers to improve accuracy and capture contextual and emotional dependencies within the text.
- Model Optimization: Hyperparameter tuning was performed to enhance the performance of each model before evaluation on the test set.

Model Training

- The models were trained on a machine with the following hardware specifications:
- Processor: Intel i5-8265U

- RAM: 8 GB
- GPU: Intel UHD Graphics 620

Metrics for Evaluation

- The primary evaluation metric used is the F1 Score, which balances precision and recall, which makes it ideal for dealing with class imbalances in emotion classification. Additional metrics, such as accuracy, precision, and recall, were also tracked to provide a comprehensive understanding of model performance.

Chapter 5

RESULT ANALYSIS AND DISCUSSION

5.1 Output

Models	F1-Score	Accuracy	Precision	Recall
Logistic Regression	0.36	0.48	0.35	0.49
Transformer (Bert)	0.45	0.51	0.43	0.51
DNN	0.44	0.30	0.45	0.31
DistilBERT VC (Text only)	0.62	0.66	0.86	0.56
DistilBERT VC (Text + Emotions)	0.65	0.70	0.87	0.59

Table 5.1 Comparison of Model Metrics

Table 5.1 summarizes the performance metrics for different methods used. Logistic Regression was able to achieve an accuracy of 0.48, F1-score of 0.36, precision of 0.35, and Recall of 0.49. Transformer (Bert) – Logistic Regression applied to BERT embeddings obtained from BERT-base-uncased model. The model was able to achieve an accuracy of 0.51, F1-score of 0.45, precision of 0.43, and Recall of 0.51. Deep Neural Network (DNN) model was able to achieve an accuracy of 0.30, F1-score of 0.44, Precision of 0.45, and Recall of 0.31. A DistilBERT VC (Text only) model which used only sentences was able to achieve an accuracy of 0.66, F1-score of 0.62, Precision of 0.86, and Recall of 0.56. DistilBERT VC (Text + Emotions) which used sentences embedded with emotion values was able to achieve an accuracy of 0.70, F1-score of 0.65, Precision of 0.87, and Recall of 0.59.

5.2 Results

Initially, using logistic regression on pre-processed data gave an F1-score of 0.36 and accuracy of 0.48 establishing a baseline. Bert-embeddings obtained from Bert-base-uncased model were then applied to Logistic Regression, which improved performance to an F1-score of 0.45 and accuracy of 0.51. To further enhance the results a Deep Neural Network (DNN) model was also explored which yielded an F1-score of 0.44 and accuracy of 0.30. To explore a different approach, DistilBERT VC models was trained on the Semeval dataset to predict human values. Model 1 used on only sentence text from semeval and serve as baseline. It focused on predicting the 38 human value labels (attained + constrained) and achieved an f1 score of 0.62 and accuracy of 0.66. Model 2 combined both sentences and it's predicted emotion labels as input, and outperformed model 1 with an F1 score of 0.65 and accuracy of 0.70. These results indicate that traditional models like Logistic Regression are simple and computationally efficient, but they demonstrate limited performance when used alone. Applying Bert embeddings significantly improves performance by capturing deeper contextual and semantic relationships within text. Combining Bert embeddings with deep learning models like DNN further boosts accuracy and effectiveness. In the value classification task, using only text worked well, but adding predicted emotions to the input further improve results. This confirms that emotional context plays a key role in improving the detection of human values from text. Overall, the approach maintains a strong balance between, semantic understanding, Computational efficiency, and predictive accuracy. This makes it ideal for complex tasks such as multi-label emotion and value classification.

Chapter 6

CONCLUSION AND FUTURE ENHANCEMENTS

This study shows the effectiveness of machine learning and deep learning methods, particularly transformer-based models, for detecting emotions in text task essential for understanding the emotional values that shape human values. The research shows that the value classifier and Bert-based models outperforms the baseline logistic regression model with a significant improvement in F1 score and overall classification metrics. Specifically, the logistic regression model achieved an F1 score of 0.36 and an accuracy of 0.48, while the BERT-based model, using embeddings generated from the Bert-base-uncased model followed by logistic regression, resulted in an improved F1 score of 0.45 and accuracy of 0.50. Further Deep Neural Network (DNN) by f1-score of 0.44 and accuracy of 0.30 and DistilBERT VC (Text Only) gave an accuracy of 0.66 and F1-score of 0.62 and DistilBERT VC (Text + Emotions) gave accuracy of 0.70 and F1-score of 0.65 which outperforms logistic regression. This improvement demonstrates the DistilBERT model's superior capability in capturing contextual dependencies and subtle emotional distinctions within text, aligning with expectations for transformer-based architectures.

While the baseline was set by logistic regression model, its limitations highlighted the necessity of a more complex model for capturing the intricacies of emotional language. The success of the DistilBERT model in handling overlapping emotional characteristics demonstrates the value of deep learning techniques in emotion-focused natural language processing (NLP) tasks.

The research concludes that the transformer based pre-trained model can be more effective in emotion detection rather the traditional machine leaning models. The fine-tuned BERT model was able to achieve better metrices in capturing emotions in comparison to directly using logistic regression on the data. Further enhancement was observed by Deep Neural Network (DNN). The study also explored the role of emotions in value classification. Models developed using the merged dataset and the semeval dataset proved to be better performer compared to other approaches. This research contributes to the field of NLP by providing a viable framework for emotion detection that can be adapted for tasks in sentiment analysis, opinion mining, and human-computer interaction studies.

The importance of this research lies in its applicability for both researchers and practitioners. For researchers, it shows the understanding of emotion detection capabilities and sets a baseline for comparing various machine learning models. For practitioners, particularly those working on applications in customer sentiment analysis, mental health diagnostics, and educational technology, the findings provide insights into selecting and implementing effective models for emotional analysis, enabling them to make data-driven decisions in real-world applications.

For future enhancements, the system can be extended to support multilingual datasets, enabling cross-cultural emotion and value analysis. Additionally, real-time applications such as chatbots can be developed to enhance user experience by recognizing emotions through text, helping users address their problems more effectively and empathetically.

For practitioners, implementing emotion detection models based on the findings of this research could improve the user experience in applications requiring sensitive engagement, such as in customer service chatbots or educational tools. Furthermore, refining the system based on the specific needs of a given application can maximize both accuracy and efficiency.

In conclusion, this study has illustrated the potential of deep learning for capturing complex human emotions in textual data, providing a foundation for further exploration into the nuanced domain of emotion and value-driven analyses. As machine learning techniques continue to evolve, their application in understanding human emotional expression promises to yield valuable insights for both academic research and practical applications. This thesis underscores the transformative role of AI in bridging the gap between data and emotional intelligence, paving the way for future innovations in this dynamic field.

REFERENCES

[1] J. Kiesel, M. Alshomary, N. Mirzakhmedova, M. Heinrich, N. Handke, H. Wachsmuth and B. Stein, "Semeval-2023 task 4: Valueeval: Identification of human values behind arguments," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.

[2] J. Kiesel, M. Alshomary, N. Handke, X. Cai, H. Wachsmuth and B. Stein, "Identifying the human values behind arguments," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

[3] G. Papadopoulos, M. Kokol, M. Dagioglou and G. Petasis, "Andronicus of rhodes at SemEval-2023 task 4: Transformer-based human value detection using four different neural network architectures," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.

[4] L. Ma, Z. Sun, J. Jiang and X. Li, "PAI at SemEval-2023 Task 4: A general multi-label classification system with class-balanced loss function and ensemble module," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.

[5] H. H. Hemati, S. H. Alavian, H. Sameti and H. Beigy, "Sutnlp at semeval-2023 task 4: Lg-transformer for human value detection," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.

[6] E. M. Monazzah and S. Eetemadi, "Prodicus at SemEval-2023 Task 4: Enhancing Human Value Detection with Data Augmentation and Fine-Tuned Language Models," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.

[7] A. R. Jafari, P. Rajapaksha, R. Farahbakhsh, G. Li and N. Crespi, "Fine-grained emotion knowledge extraction in human values: An interdisciplinary analysis," in *The 12th International Conference on Complex Networks and their Applications*, 2023.

[8] S. Saha and R. Srihari, "Rudolf christoph eucken at semeval-2023 task 4: An ensemble approach for identifying human values from arguments," *arXiv preprint arXiv:2305.05335,* 2023.

[9] D. Zaikis, S. D. Stefanidis, K. Anagnostopoulos and I. Vlahavas, "Aristoxenus at SemEval-2023 Task 4: A Domain-Adapted Ensemble Approach to the Identification of Human Values behind Arguments," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023.

[10] K. Machová, M. Szabóova, J. Paralič and J. Mičko, "Detection of emotion by text analysis using machine learning," *Frontiers in Psychology,* vol. 14, p. 1190326, 2023.

[11] J. J. Deng, C. H. C. Leung, P. Mengoni and Y. Li, "Emotion recognition from human behaviors using attention model," in *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2018.

[12] A. R. Jafari, P. Rajapaksha, R. Farahbakhsh, G. Li and N. Crespi, "Unveiling Human Values: Analyzing Emotions behind Arguments," *Entropy,* vol. 26, p. 327, 2024.

[13] A. Bandhakavi, N. Wiratunga, D. Padmanabhan and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern recognition letters,* vol. 93, p. 133–142, 2017.

[14] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *Journal of computational science,* vol. 36, p. 101003, 2019.

[15] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[16] K. R. Scherer and H. G. Wallbott, *ISEAR Dataset on Hugging Face,* 2022.

[17] Y. Li, H. Su, X. Shen, W. Li, Z. Cao and S. Niu, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," in *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*, 2017.

[18] S. Bird, E. Klein and E. Loper, Natural Language Processing with Python, O'Reilly Media, Inc., 2009.

[19] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

[20] I. Khawaja, *The Schwartz Theory of Basic Values, and Some Implications for Political Philosophy,* 2015.

[21] M. Alshomary, R. E. Baff, T. Gurcke and H. Wachsmuth, "The moral debater: A study on the computational generation of morally framed arguments," *arXiv preprint arXiv:2203.14563,* 2022.

[22] Y. Ajjour, M. Alshomary, H. Wachsmuth and B. Stein, "Modeling frames in argumentation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[23] T. J. M. Bench-Capon, "Persuasion in practical argument using value-based argumentation frameworks," *Journal of Logic and Computation,* vol. 13, p. 429–448, 2003.

[24] G. Feldman, "Personal values and moral foundations: Examining relations and joint prediction of moral variables," *Social Psychological and Personality Science,* vol. 12, p. 676–686, 2021.

[25] R. M. Entman, "Framing: Toward clarification of a fractured paradigm," *Journal of communication,* vol. 43, p. 51–58, 1993.

Detection of ethical principles based on argument using pre-Trained models

ORIGINALITY REPORT

12%	11%	11%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to VIT University Student Paper	1%
2	aclanthology.org Internet Source	1%
3	www.coursehero.com Internet Source	1%
4	www.mdpi.com Internet Source	1%
5	bitswgl.ac.in Internet Source	1%
6	"ECAI 2020", IOS Press, 2020 Publication	<1%
7	Submitted to Liverpool John Moores University Student Paper	<1%
8	H L Gururaj, Francesco Flammini, V Ravi Kumar, N S Prema. "Recent Trends in Healthcare Innovation", CRC Press, 2025 Publication	<1%
9	Parikshit N. Mahalle, Namrata N. Wasatkar, Gitanjali R. Shinde. "Data-Centric Artificial Intelligence for Multidisciplinary Applications", CRC Press, 2024 Publication	<1%

10	scholarshare.temple.edu Internet Source	<1%
11	Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Artificial Intelligence, Blockchain, Computing and Security", CRC Press, 2023 Publication	<1%
12	Debasis Chaudhuri, Jan Harm C Pretorius, Debashis Das, Sauvik Bal. "International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023) - Proceedings of the International Conference on Security, Surveillance and Artificial Intelligence (ICSSAI-2023), Dec 1–2, 2023, Kolkata, India", CRC Press, 2024 Publication	<1%
13	ceur-ws.org Internet Source	<1%
14	paperswithcode.com Internet Source	<1%
15	Jang, Youngsun. "Optimizing Large Language Models and Multimodal Approaches for Biomedical Publication and Satellite Imagery", South Dakota State University, 2024 Publication	<1%
16	Thompson Stephan. "Artificial Intelligence in Medicine", CRC Press, 2024 Publication	<1%
17	sciencecast.org Internet Source	<1%
18	arxiv.org Internet Source	<1%

19	Akter, Shapna. "Integrated Machine Learning Framework for Mitigation of Buffer Overflow, Software Supply Chain, and Adversarial Attacks", The University of West Florida, 2024 Publication	<1 %
20	"Document Analysis and Recognition – ICDAR 2021", Springer Science and Business Media LLC, 2021 Publication	<1 %
21	Kashfia Sailunaz, Reda Alhajj. "Emotion and sentiment analysis from Twitter text", Journal of Computational Science, 2019 Publication	<1 %
22	Submitted to National College of Ireland Student Paper	<1 %
23	Thangavel Murugan, W. Jai Singh. "Cybersecurity and Data Science Innovations for Sustainable Development of HEICC - Healthcare, Education, Industry, Cities, and Communities", CRC Press, 2025 Publication	<1 %
24	www2.mdpi.com Internet Source	<1 %
25	dspace.vut.cz Internet Source	<1 %
26	"Value Engineering in Artificial Intelligence", Springer Science and Business Media LLC, 2024 Publication	<1 %
27	Anh Ngo, Jan Kocoń. "Integrating personalized and contextual information in fine-grained emotion recognition in text: A multi-source	<1 %

fusion approach with explainability", Information Fusion, 2025
Publication

28	Pijush Samui, Sanjiban Sekhar Roy, Wengang Zhang, Y H Taguchi. "Machine Learning and IoT Applications for Health Informatics", CRC Press, 2024 Publication	<1 %
29	www.nature.com Internet Source	<1 %
30	www.researchgate.net Internet Source	<1 %
31	www.tfzr.rs Internet Source	<1 %
32	Ahmed, Nourin. "Optimizing Longitudinal Data Representation to Predict Clinical Scores Using an End-to-End Machine Learning Pipeline: A Case Study in Parkinson's Disease.", University of Windsor (Canada) Publication	<1 %
33	Chris J. Vargo. "The Computational Content Analyst - Using Machine Learning to Classify Media Messages", Routledge, 2024 Publication	<1 %
34	Sakshi Gupta, Umesh Gupta, Moolchand Sharma, Kamal Malik. "Generative Intelligence in Healthcare - Transforming Patient Care with AI Creativity", CRC Press, 2025 Publication	<1 %
35	Submitted to Zikura International College Student Paper	<1 %
36	scholar.uoc.ac.in Internet Source	<1 %

37	T. Mariprasath, Kumar Reddy Cheepati, Marco Rivera. "Practical Guide to Machine Learning, NLP, and Generative AI: Libraries, Algorithms, and Applications", River Publishers, 2024 Publication	<1 %
38	"Natural Language Processing and Chinese Computing", Springer Science and Business Media LLC, 2019 Publication	<1 %
39	Submitted to University of Hertfordshire Student Paper	<1 %
40	Submitted to University of Stirling Student Paper	<1 %
41	export.arxiv.org Internet Source	<1 %
42	Submitted to Aston University Student Paper	<1 %
43	Submitted to Taylor's Education Group Student Paper	<1 %
44	Submitted to Universidad Francisco de Vitoria Student Paper	<1 %
45	Vasilakes, Jake A.. "Factualizing Biomedical Event Extraction", The University of Manchester (United Kingdom) Publication	<1 %
46	epdf.pub Internet Source	<1 %
47	munin.uit.no Internet Source	<1 %
48	pure.uva.nl Internet Source	<1 %

49	eprints.utm.my Internet Source	<1 %
50	ojs.unud.ac.id Internet Source	<1 %
51	www.gamerzplanet.net Internet Source	<1 %
52	www.ijisae.org Internet Source	<1 %
53	www.ijsr.net Internet Source	<1 %
54	Yu Su, Xiangyu Zheng, Junyu Lu, Yi Gong, Qijuan Gao, Shuanghong Shen, Qi Liu. "Semantic distillation and enhanced diagnostic alignment: A novel approach for depression detection in social media", Expert Systems with Applications, 2025 Publication	<1 %
55	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
56	"HHAI 2023: Augmenting Human Intellect", IOS Press, 2023 Publication	<1 %
57	Joe D. Menke, Halil Kilicoglu, Neil R. Smalheiser. "Publication Type Tagging using Transformer Models and Multi-Label Classification", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
58	Kristi Rüütel, R. David Parker, Liilia Lõhmus, Anti Valk, Toivo Aavik. "HIV and STI Testing and Related Factors Among Men Who Have	<1 %

Sex with Men in Estonia", AIDS and Behavior, 2016
Publication

59	Senemmar, Soroush. "Shipboard Power Systems Load Monitoring and Energy Management for Resilience Enhancement", The University of Texas at Dallas Publication	<1 %
60	Zhaozhen Xu, Philippa Davies, Louise AC Millard, Lam Teng et al. "M-PreSS: A Model Pre-training Approach for Study Screening in Systematic Reviews", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
61	api.repository.cam.ac.uk Internet Source	<1 %
62	nika.vin.ua Internet Source	<1 %
63	"Advanced Concepts for Intelligent Vision Systems", Springer Science and Business Media LLC, 2023 Publication	<1 %
64	"Artificial Neural Networks and Machine Learning – ICANN 2021", Springer Science and Business Media LLC, 2021 Publication	<1 %
65	Jyotika Singh. "Natural Language Processing in the Real World - Text Processing, Analytics, and Classification", CRC Press, 2023 Publication	<1 %
66	Konstantinos Bacharidis, Antonis Argyros. "Improving Deep Learning Approaches for Human Activity Recognition based on Natural	<1 %

Language Processing of Action Labels", 2020 International Joint Conference on Neural Networks (IJCNN), 2020
Publication

67	Nourah Alswaidan, Mohamed El Bachir Menai. "A survey of state-of-the-art approaches for emotion recognition in text", Knowledge and Information Systems, 2020 Publication	<1 %
68	Zhou Junjie, Paolo Mengoni. "Spot Gold Price Prediction Using Financial News Sentiment Analysis", 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2020 Publication	<1 %
69	core.ac.uk Internet Source	<1 %
70	docplayer.net Internet Source	<1 %
71	dspace02.jaist.ac.jp Internet Source	<1 %
72	flore.unifi.it Internet Source	<1 %
73	forum.tomshw.it Internet Source	<1 %
74	iieta.org Internet Source	<1 %
75	ijsrem.com Internet Source	<1 %
76	repository.library.carleton.ca Internet Source	<1 %

77	rosap.ntl.bts.gov Internet Source	<1 %
78	shmpublisher.com Internet Source	<1 %
79	www.acadlore.com Internet Source	<1 %
80	www.brainwareuniversity.ac.in Internet Source	<1 %
81	www.csitcp.com Internet Source	<1 %
82	www.frontiersin.org Internet Source	<1 %
83	www.ixa.eus Internet Source	<1 %
84	www.ncbi.nlm.nih.gov Internet Source	<1 %
85	"Value Engineering in Artificial Intelligence", Springer Science and Business Media LLC, 2025 Publication	<1 %
86	Amjad Anvari-Moghaddam, Sina Ghaemi, Shi You, Frede Blaabjerg. "Power-to-X in Regional Energy Systems - Planning, Operation, Control, and Market Perspectives", CRC Press, 2025 Publication	<1 %
87	"Computer Applications", Springer Science and Business Media LLC, 2024 Publication	<1 %
88	Mohamed Lahby, Al-Sakib Khan Pathan, Yassine Maleh. "Combatting Cyberbullying in	<1 %

Digital Media with Artificial Intelligence", CRC Press, 2023
Publication

89	Toliyat, Amir. "Analysis of the Inter-Annotators Agreement and Its Effect on the System of Anti-Asian Hate Crime Detection on Twitter During COVID-19", City University of New York Publication	<1 %

Exclude quotes	On	Exclude matches	< 3 words
Exclude bibliography	On		