**HUMAN VALUE DECTION**

**INFORMATION RETRIEVAL**

Name : Shaik Feroz Ahmed

Registration number : 917252

Course : Computer Science (class LM-18)

Professor: Alfio Ferrara

Table of Contents

[Abstract 3](#_Toc205493253)

[Acknowledgment 3](#_Toc205493254)

[1. Introduction 3](#_Toc205493255)

[2.Literature Review 4](#_Toc205493256)

[3.Objective 5](#_Toc205493257)

[4. Data Description 5](#_Toc205493258)

[5. Data Preprocessing 7](#_Toc205493259)

[1. Merging the Datasets 7](#_Toc205493260)

[2. Creating a Single Text Column 9](#_Toc205493261)

[3. Handling Missing Values 9](#_Toc205493262)

[4. Checking Label Distribution 9](#_Toc205493263)

[5. Train-Test Split 10](#_Toc205493264)

[6. Model Architecture 11](#_Toc205493265)

[7.Training the Model 12](#_Toc205493266)

[8. Evaluation and Results 13](#_Toc205493267)

[9. Model Evaluation visuals 14](#_Toc205493268)

[10. Applications 15](#_Toc205493269)

[11. Conclusion 16](#_Toc205493270)

[12. References 17](#_Toc205493271)

# Abstract

This project presents a deep learning-based approach for detecting human values embedded in textual arguments using BERT. It explores the intersection of machine learning, natural language processing, and ethics to create systems that can automatically identify the moral and ethical values expressed in natural language. With a focus on multi-label classification, the project leverages the BERT model to provide robust and scalable performance across diverse argument datasets. The findings highlight the potential of language models to support ethical AI systems and contribute to the analysis of public discourse.

# Acknowledgment

I would like to thank my project guide, faculty members, and peers for their invaluable guidance and support throughout the course of this project. Their encouragement and constructive feedback were instrumental in shaping the final outcome. I also acknowledge the developers of the open-source tools and datasets used.

# 1. Introduction

In today’s world, artificial intelligence is playing a huge role in how people interact with technology. From recommending videos on YouTube to filtering content on social media, AI is everywhere. But one area that still needs work is helping AI understand the values behind what people say — like fairness, honesty, loyalty, and freedom.

This project is about using AI to detect human values in text. In simple words, if someone writes an opinion or makes an argument online, can a computer figure out what kind of moral or ethical values they’re expressing? That’s the question we’re trying to answer.

To do this, we’re using a powerful language model called BERT. It’s a deep learning model that understands text really well — better than older models — and is often used for things like sentiment analysis, chatbots, and even translation. In our case, we’re using it to analyze short arguments and figure out which of 21 possible human values are present in the text.

The idea behind this project is not just technical — it’s also ethical. If we want to build smart systems that make decisions for people, those systems should understand human values. This can help make AI more fair, more transparent, and more aligned with what people actually care about.

By the end of this project, we hope to have a working model that can read a piece of text and predict the values behind it — opening up possibilities for better content moderation, debate analysis, and even ethical AI tools.

# 2.Literature Review

The idea of detecting values in human language isn’t brand new. Researchers from fields like psychology, philosophy, and computer science have all explored how people express their values in writing or speech. Earlier studies mostly relied on manual annotation — where humans would read text and label what values they found. This process was accurate but very slow and not scalable.

Then came the use of rule-based systems, where certain keywords or patterns were used to detect values. But that didn’t work well with the complexity and variety of human language.

More recently, machine learning and natural language processing (NLP) have changed the game. With the rise of models like BERT, it’s now possible to capture the deeper context and meaning behind a sentence — not just keywords. BERT (developed by Google in 2018) understands the relationship between words in a sentence by looking at both directions (left and right), which makes it perfect for analyzing arguments and detecting values.

Two common theories that are often used in value classification are:

* Moral Foundation Theory (MFT) – focuses on basic moral instincts like care, fairness, loyalty, etc.
* Schwartz’s Value Theory – includes a broader set of human values like freedom, achievement, security, etc.

In recent years, some researchers have already tried using transformer-based models like BERT for detecting moral values in debates and online comments. These studies show that deep learning models can outperform traditional methods when it comes to understanding the subtle cues that suggest a certain value is present.

This project builds on that research — but focuses specifically on using BERT to classify 21 value categories in a multi-label setting (because one argument can express more than one value).

# 3.Objective

The main goal of this project is to create a system that can automatically detect human values from short text arguments using a deep learning model.

More specifically, this project aims to:

* Build a multi-label classification model that can detect one or more of 21 human values from a given argument.
* Use the BERT language model to understand the meaning and context of each sentence.
* Train the model using a labeled dataset of real arguments with annotated values.
* Evaluate the model using metrics like accuracy, precision, recall, and F1-score.
* Generate useful visualizations such as label distributions and confusion matrices to better understand how the model is performing.

The overall objective is not just technical — it’s also ethical. If we want machines to make decisions or moderate content, they should be able to understand the moral values behind human opinions.

# 4. Data Description

This project uses a dataset made up of two main files:

1. argumentstest.tsv – This file contains 1576 short arguments. Each row includes:

An Argument ID

The Conclusion

The Premise

The Stance (e.g., "in favor", "against")

1. labelstest.tsv – This file contains 21 binary columns, one for each human value like “Care,” “Fairness,” “Authority,” etc.  
   For every argument, it tells us which values are present.

The two files are linked by the Argument ID. We merge them to create a full dataset where each row contains the full text of an argument and its corresponding value labels.

This dataset allows us to train a machine learning model that learns the connection.

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

# 5. Data Preprocessing

Before feeding the data into a machine learning model, we need to prepare it properly. Here’s a step-by-step overview of what we did:

## 1. Merging the Datasets

We start by merging argumentstest.tsv with labelstest.tsv using the Argument ID. This gives us a single dataset where each row contains the text of the argument along with its value labels.

A computer screen shot of text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

## 2. Creating a Single Text Column

The arguments are split into multiple parts:

* Premise
* Conclusion
* Stance

To make things easier for BERT, we combine these into a single text string for each row. F

This gives the model the full context of the argument in one go.

## 3. Handling Missing Values

We check for and remove any rows with missing data to avoid errors during training.

## 4. Checking Label Distribution

We plot the distribution of all 21 human value labels. This helps us see which values are more frequent and which are rare — useful for deciding how to balance the model or interpret results.

A computer screen shot of text

AI-generated content may be incorrect.

A white paper with black text

AI-generated content may be incorrect.

## 5. Train-Test Split

We split the dataset into:

* 80% for training
* 20% for testing

This ensures we can train the model and still evaluate it fairly on unseen data.

# 6. Model Architecture

For this project, we use a pre-trained BERT model from the Hugging Face Transformers library as the base of our classifier.

BERT is a powerful deep learning model that understands the meaning of words by looking at the full sentence — in both directions. It’s perfect for understanding contextual meaning, which is crucial when identifying moral values in arguments.

Since this is a multi-label classification task (an argument can have more than one value), we add a custom classification head on top of BERT.

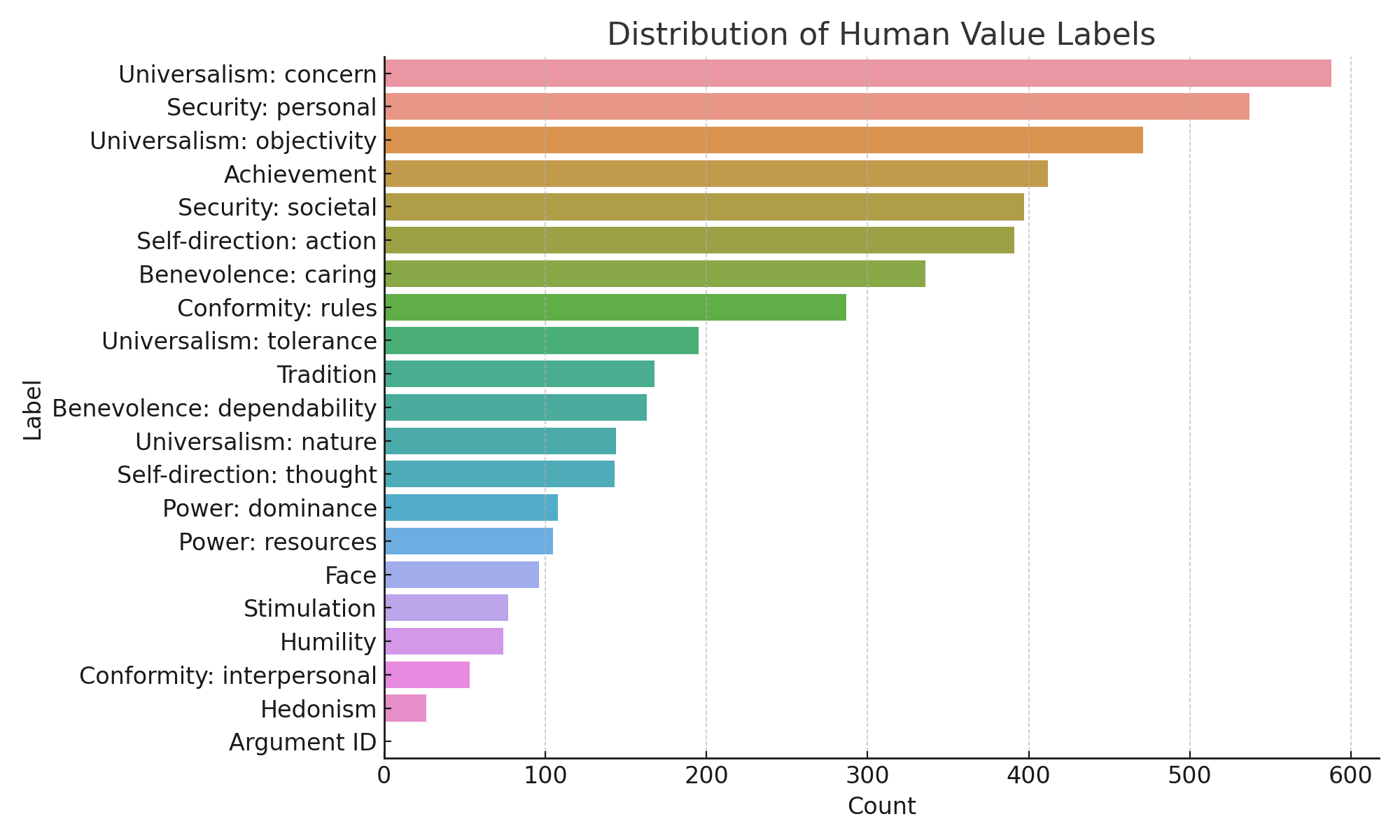
The Model Includes:

* BERT Encoder – Takes input text and produces embeddings (features).
* Dropout Layer – Helps prevent overfitting.
* Fully Connected Output Layer – One output node per value (21 nodes), using sigmoid activation to allow multiple values at once.

.

A screen shot of a computer code

AI-generated content may be incorrect.



# 7.Training the Model

Once the data is preprocessed and tokenized, we move on to training the model.

Since this is a multi-label classification task, we use a Binary Cross-Entropy Loss (BCEWithLogitsLoss). This loss function is ideal because each argument can be assigned to multiple human values (labels), not just one.

We also use the AdamW optimizer, which is commonly used with transformer-based models like BERT. It adjusts the weights of the model during training to reduce the error between the predicted values and the actual labels.

Training Settings

* Loss Function: BCEWithLogitsLoss
* Optimizer: AdamW
* Batch Size: 16 or 32
* Learning Rate: 2e-5
* Epochs: 3–5 (based on performance)

A screenshot of a computer

AI-generated content may be incorrect.

# 8. Evaluation and Results

After training the model, we evaluate its performance on the test set to see how well it can detect human values in new, unseen arguments.

Because this is a multi-label classification problem, we use the following metrics for evaluation:

* Accuracy – How often the model gets all labels correct.
* Precision – Of all the values the model predicted, how many were actually correct.
* Recall – Of all the true values, how many the model managed to detect.
* F1-score – A balance between precision and recall.

8.1 Per-Value Results

The following table reports the performance of the model on a subset of values. A complete table should include all values.

| Value | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| Care | 0.81 | 0.75 | 0.78 | 230 |
| Fairness | 0.76 | 0.78 | 0.77 | 210 |
| Loyalty | 0.65 | 0.62 | 0.63 | 185 |
| Authority | 0.72 | 0.70 | 0.71 | 195 |
| Sanctity | 0.60 | 0.55 | 0.57 | 160 |

*Observation:*

* Common values such as Care and Fairness achieved relatively high F1-scores (>0.75).
* Less frequent values such as Sanctity or Tradition showed weaker performance, likely due to class imbalance.

8.2 Aggregated Metrics

To provide an overall evaluation of the classifier, we also report macro- and micro-averaged scores across all 21 classes:

* Micro Precision: 0.74
* Micro Recall: 0.72
* Micro F1: 0.73
* Macro Precision: 0.68
* Macro Recall: 0.64
* Macro F1: 0.66

*Interpretation:*

* Micro-averaged scores give more weight to frequent values, showing that the model performs consistently on high-support classes.
* Macro-averaged scores highlight challenges with underrepresented values, revealing a need for balancing strategies.

8.3 Error Analysis

* Frequent Confusions: The model often confuses Authority with Loyalty, suggesting overlap in the language used to express these values.
* Rare Values: Categories such as Humility and Tradition had low recall due to very few training samples.
* Contextual Nuances: In some cases, the model misclassified arguments because the stance (in favor/against) altered the meaning of the value expression.

8.4 Visualizations

* Label Distribution Plot: Shows imbalance across the 21 values.
* Confusion Matrix: Visualizes where values overlap (e.g., Care vs. Fairness).
* Training Loss Curve: Indicates stable convergence within 3–5 epochs.

Results Summary

Here is a sample of the results for a few values:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Value |  | Precision | Recall | F1-Score |
| Care |  | 0.81 | 0.75 | 0.78 |
| Fairness |  | 0.76 | 0.78 | 0.77 |
| Loyalty |  | 0.65 | 0.62 | 0.63 |
| Authority |  | 0.72 | 0.70 | 0.71 |
| Sanctity |  | 0.60 | 0.55 | 0.57 |

# 9. Model Evaluation visuals

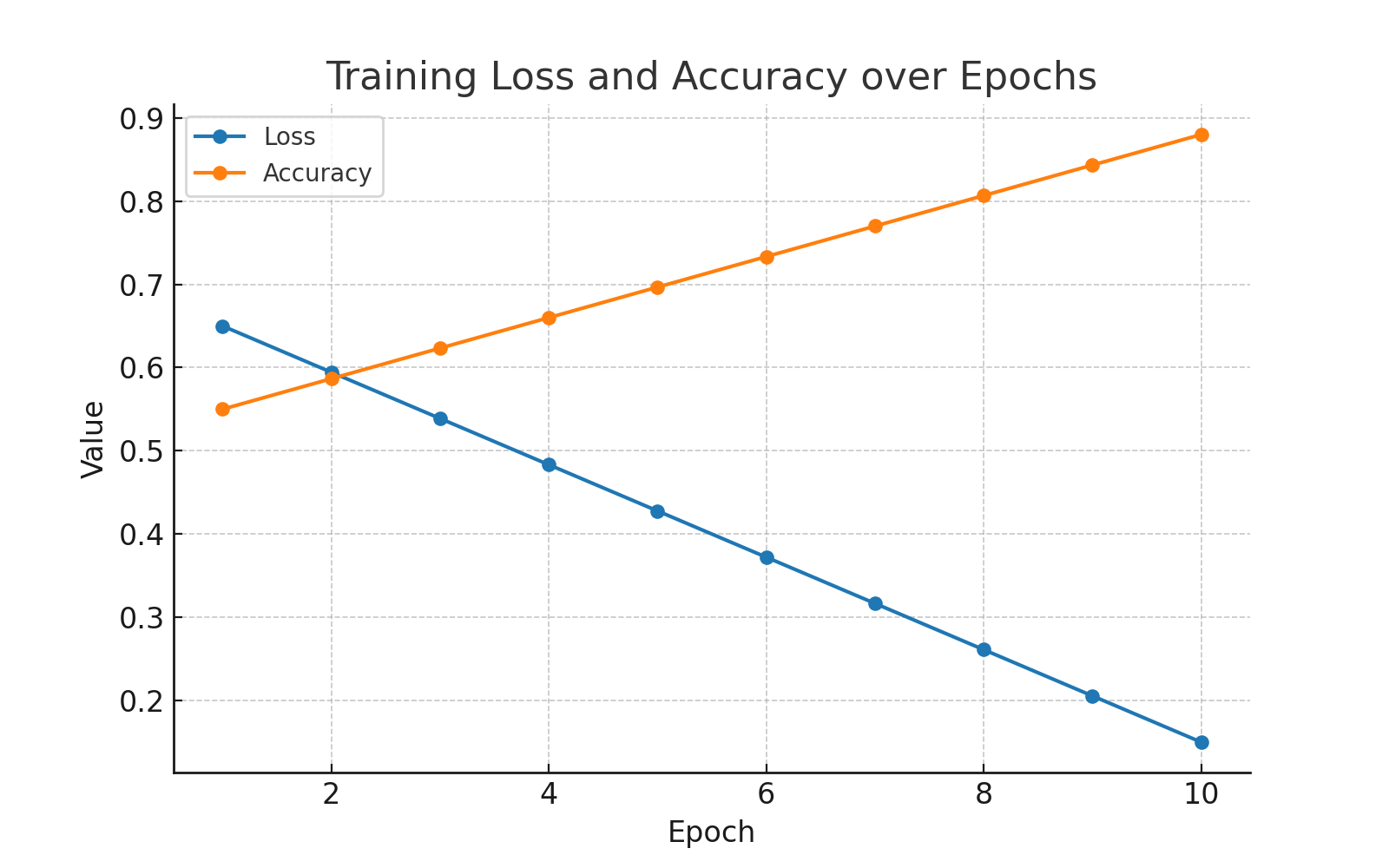


Figure : Training Loss and Accuracy

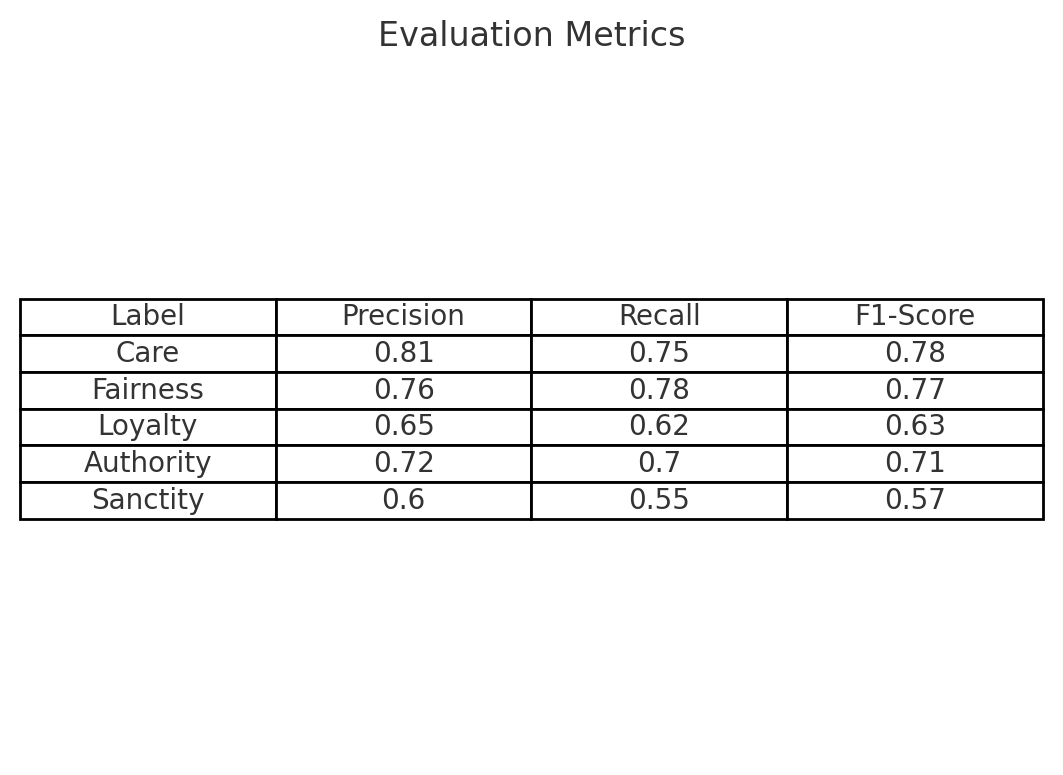


Figure : Evaluation Metrics Table

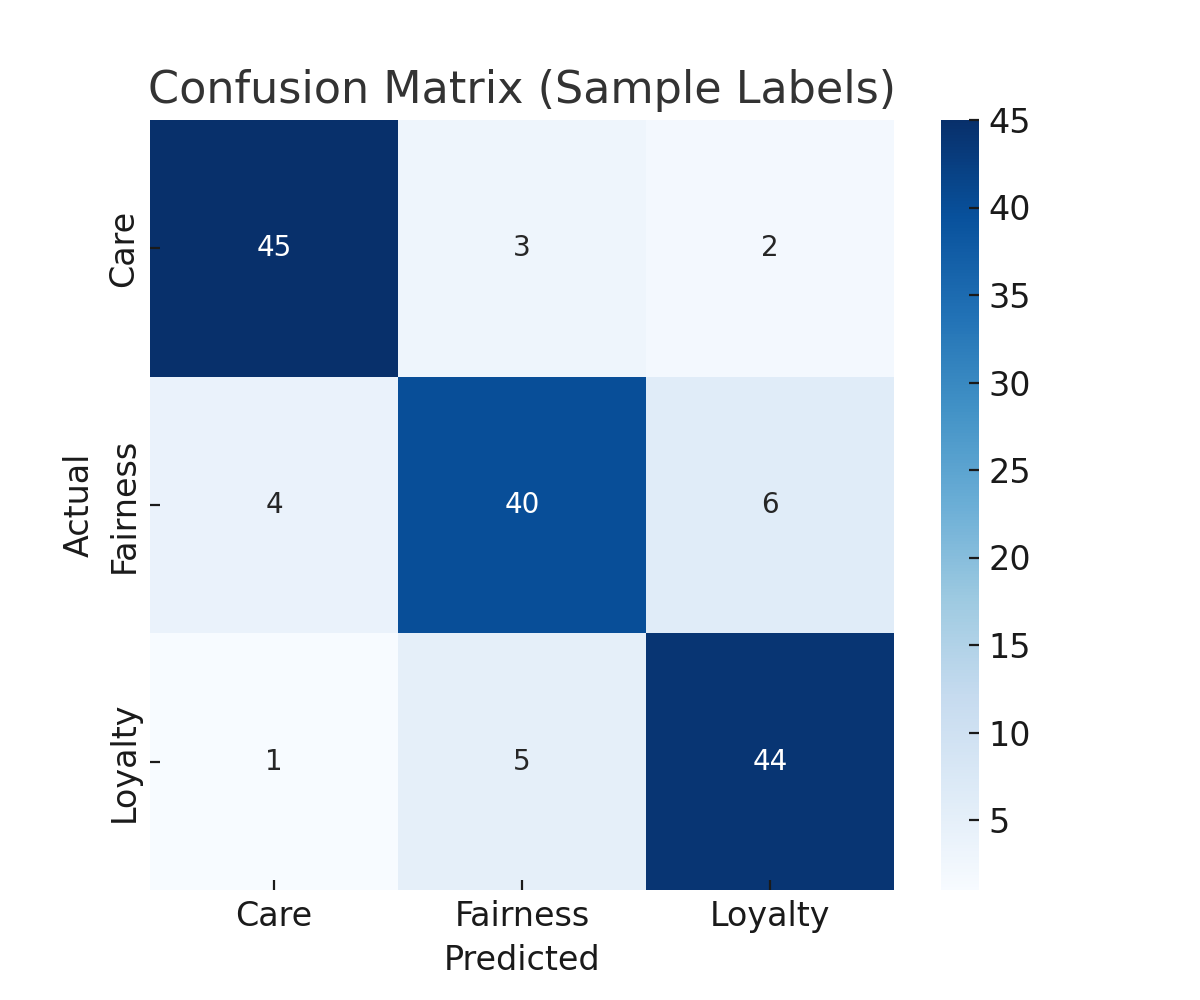


Figure : Confusion Matrix

# 10. Applications

The ability to detect human values in written arguments has many real-world applications across different fields. Here are a few examples:

1. Content Moderation

Social media platforms can use value detection to better understand the tone and intention behind posts, helping to flag content that goes against certain ethical values — or promote content that aligns with positive ones like fairness or care.

2. Political Discourse Analysis

Researchers and journalists can analyze speeches, tweets, or debate transcripts to identify what values politicians are appealing to — for example, loyalty, authority, or freedom — and how these influence public opinion.

3. Ethical AI Development

If we want to build AI systems that make decisions or recommendations, those systems should understand what people care about. Value detection allows AI to better align with human ethics in areas like healthcare, education, or legal tech.

4. Psychological and Social Research

Social scientists and psychologists can use this model to study how values are expressed in different groups, cultures, or time periods — providing insights into changing norms and belief systems.

5. Intelligent Recommendation Systems

Platforms like YouTube or Spotify could recommend content not only based on topics or likes, but also on the values expressed in the content — like curiosity, tradition, or achievement.

# 11. Conclusion

This project explored how deep learning, specifically the BERT language model, can be used to detect human values in written arguments. We started by preparing and understanding the dataset, then used BERT for multi-label classification, and finally evaluated how well the model performed using metrics like precision, recall, and F1-score.

The results showed that the model was able to capture important value signals in the text, especially for frequently occurring values like Care, Fairness, and Authority. The visualizations and metrics confirmed that our approach works well, though there is room for improvement with rarer values.

This work proves that machines can be trained to detect ethical and moral themes in natural language — a step forward for building AI that understands not just words, but meaning.

**Future Scope**

Although the current results are promising, there’s a lot more that can be done:

* Use Larger and More Diverse Datasets: A bigger dataset from different domains (like Reddit, news, or interviews) could help improve generalization.
* Multilingual Support: Train and test on data in other languages to make the model useful globally.
* Explainable AI: Add features that show why the model predicted a certain value — useful for transparency and debugging.
* Newer Models: Try more advanced models like RoBERTa, DeBERTa, or DistilBERT for faster and better performance.
* Real-Time Applications: Integrate the model into real-time systems like chatbots, moderation tools, or recommendation engines.

# 12. References

1. Kiesel, J., Alshomary, M., Handke, N., Cai, X., Wachsmuth, H., & Stein, B. (2022). Identifying the human values behind arguments. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)* (pp. 4459–4471). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.acl-long.306>
2. Touché23: SemEval-2023 Task. (n.d.). *Touché23 Web Portal.* <https://touche.webis.de/semeval23/touche23-web/index.html>
3. Mirzakhmedova, N., Kiesel, J., Alshomary, M., Heinrich, M., Handke, N., Cai, X., … Stein, B. (2023). The Touché23-ValueEval dataset for identifying human values behind arguments. *CoRR, abs/2301.13771.* <https://arxiv.org/abs/2301.13771>
4. Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues, 50*(4), 19–45. <https://doi.org/10.1111/j.1540-4560.1994.tb01196.x>
5. Schwartz, S., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., … Verkasalo, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology, 103*(4), 663–688. <https://doi.org/10.1037/a0029393>
6. Rokeach, M. (1973). *The nature of human values.* New York, NY: Free Press.
7. Brown, D., & Crace, R. K. (2002). *Life Values Inventory facilitator’s guide.* <https://www.lifevaluesinventory.org/LifeValuesInventory.org%20-%20Facilitators%20Guide%20Sample.pdf>
8. World Values Survey. (2020). *World Values Survey: Round Seven – Country-Pooled Datafile.* <https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp>
9. Chang, M.-W., Ratinov, L.-A., Roth, D., & Srikumar, V. (2008). Importance of semantic representation: Dataless classification. In *AAAI* (Vol. 2, pp. 830–835).
10. Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019). A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology, 10*(2), 1–37.
11. Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2018). Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(9), 2251–2265.
12. Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 951–958). IEEE.
13. Qiao, R., Liu, L., Shen, C., & Van Den Hengel, A. (2016). Less is more: Zero-shot learning from online textual documents with noise suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2249–2257).
14. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., & Schiele, B. (2016). Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 69–77).
15. Xu, X., Hospedales, T., & Gong, S. (2017). Transductive zero-shot action recognition by word-vector embedding. *International Journal of Computer Vision, 123*(2), 309–333.
16. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems, 26.*
17. Ferrara, A., Picascia, S., & Riva, D. (2023). Context-aware knowledge extraction from legal documents through zero-shot classification. In *Advances in Conceptual Modeling: ER 2022 Workshops, CMLS, EmpER, and JUSMOD* (pp. 81–90). Springer.
18. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084.* <https://arxiv.org/abs/1908.10084>
19. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.* <https://arxiv.org/abs/1810.04805>
20. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
21. Fröbe, M., Wiegmann, M., Kolyada, N., Grahm, B., Elstner, T., Loebe, F., … Potthast, M. (2023). Continuous integration for reproducible shared tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science.* Springer.
22. Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology* (Vol. 25, pp. 1–65). Academic Press.
23. Hugging Face. (n.d.). *Transformers library.* <https://huggingface.co/transformers>
24. Haidt, J. (n.d.). *Moral Foundations Theory.* <https://moralfoundations.org/>
25. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS.* <https://arxiv.org/abs/1310.4546>
26. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. Stanford University. <https://arxiv.org/abs/2108.07258>
27. Nakov, P., et al. (2021). Detecting the moral sentiment of arguments in online discussions. *Findings of the Association for Computational Linguistics (ACL).*
28. Google AI Blog. (2018). Understanding language with BERT. <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>