UNIVERSITÀ
DEGLI STUDI
DI MILANO

Information Retrieval
Human Value Detection

April 2024

Farnoush Zeidi
farnoush.zeidikolehparcheh@studenti.unimi.it
Matriculation number: 11327A
Course Assesment Project for Information Retrieval
Prof. Alfio Ferrara

Abstract

Detecting human values from text poses a significant challenge in natural language processing. It involves understanding individuals' beliefs and principles based on written text. The complexity arises from the implicit nature of contextual hints, which can be elusive to computational methods. Nevertheless, this area of study has garnered considerable interest within the natural language processing community, prompting us to conduct comparative analyses of various models.

In our project, we focus on training and evaluating different models to see how well they perform in identifying human values in textual data. Inspired by the objectives of SemEval Task 4, our goal is to compare the effectiveness of these models in recognizing specific value categories within text. Each model undergoes training on annotated data and is then assessed based on its ability to predict the presence or absence of values in given textual arguments. Through this comparative approach, we aim to gain insights into the strengths and weaknesses of different model architectures and techniques in addressing this task.

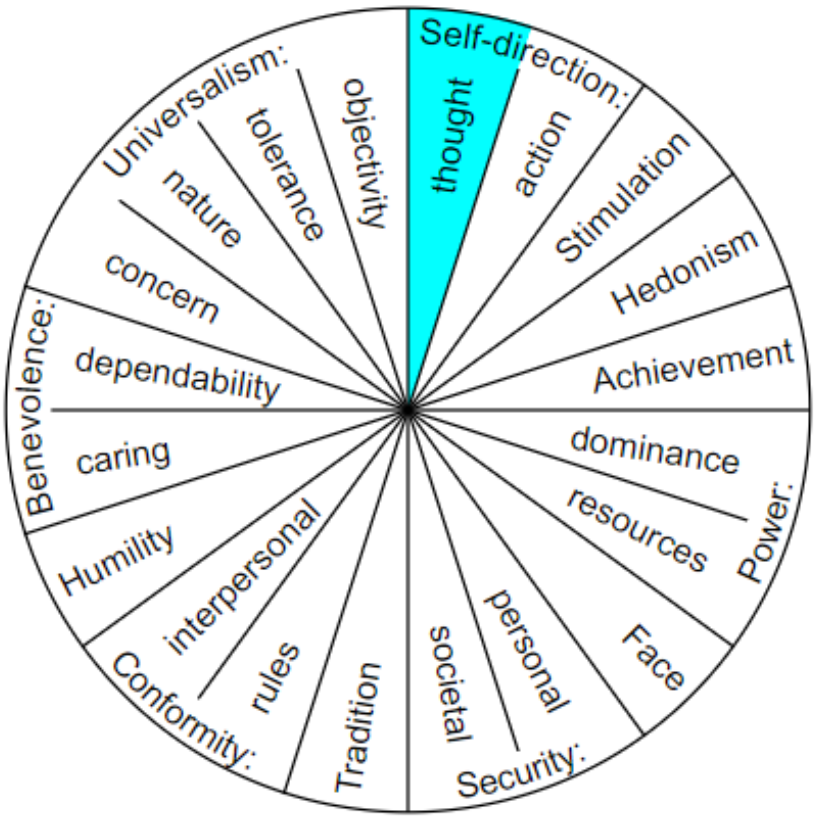
Contents

1 Introduction

Deciphering human values embedded in text represents a formidable obstacle in natural language processing. The inherent subtleties of contextual cues pose a significant challenge for computational extraction. However, this area has emerged as a focal point within the NLP community, prompting the inception of SemEval Task 4. This task addresses the complexity by framing it as a multi-label classification endeavor, aiming to forecast the presence of predetermined value categories within textual arguments. These arguments are structured into a trio of elements: a conclusion, stance, and premise, where the objective is to ascertain the existence or absence of specific values, encapsulated as a binary vector $y = [0, 1]^{20}$. Furthermore, the evaluation of this project also considers the F1 score, ensuring a comprehensive assessment of model performance.
The code for this project is available on GitHub at https://github.com/farnoush-zeidi/Info_Ret.

1.1 Datasets

This task involves utilizing a set of 20 value categories derived from the social science literature. Runs, also known as approaches or systems, can be developed to detect one, a subset, or all of these values within arguments. The dataset for this project is provided in tab-separated values files, each with a header line. The dataset includes files for training, validation, and testing.



Arguments Data

The arguments dataset (arguments-training/validation/test.tsv) contains one argument per line, consisting of its unique argument ID, the conclusion, the stance of the premise towards the conclusion, and the premise itself. For example:

Argument ID	Conclusion	Stance	Premise
A03014	Entrapment should be legalized	in favor of	anyone who commits a crime should be prosecuted
A04019	Social media brings more harm than good	in favor of	social media results in the fomo culture which is unhealthy for individuals.
A05025	Holocaust denial should be a criminal offence	against	everyone should have the right to their own opinion.
A05026	Blockade of the Gaza Strip should be ended	in favor of	blockade means occupation which is illegal and should be stopped
A05038	Assisted suicide should be a criminal offence	against	if someone is in extreme pain why make them suffer? this is not criminal
A05051	Homeopathy brings more harm than good	against	homeopathy is a natural way to cure a disease.

Labels Data

The labels dataset (labels-training/validation/test.tsv) contains one argument per line, with its unique argument ID and columns for each of the 20 value categories. Each category is represented by a binary value (1 for present, 0 for absent). For example:

Argument ID	Self-direction: thought	Self-direction: action	...	Universalism: objectivity
A03014	0	0	...	0
A04019	1	0	...	0
A05025	1	1	...	0
A05026	0	1	...	0
A05038	1	1	...	0
A05051	0	0	...	0

In addition to the primary dataset, three additional datasets are provided for evaluating model robustness:

Validation-zhihu: Derived from the recommendation and hotlist section of the Chinese question-answering website Zhihu, this dataset includes labels.
Test-nahjalbalagha: Based on and derived from the Nahj al-Balagha, labels for this dataset were kept secret during the competition.
Test-nyt: Comprised of New York Times articles related to the Coronavirus, labels for this dataset were also kept secret during the competition and may require downloading articles unless using Docker submission.

2 Preprocessing

In the data preprocessing phase, rigorous steps are employed to ensure the data's uniformity and readiness for subsequent analysis. Initially, the data and corresponding labels are merged into a single dataframe, ensuring seamless integration and coherence. Following this integration, the input data undergoes transformation, resulting in the creation of a new column termed "text." This column aggregates information from the premise, conclusion, and stance columns through concatenation, streamlining the dataset for further analysis and modeling tasks.
Additionally, another column named "category" is generated during this phase. This column serves to indicate which label is assigned a value of 1, thus providing a concise overview of the relevant category within the dataset. By including the column name in the "category" column, the dataset is enhanced with clear categorization, facilitating ease of interpretation and analysis. These systematic procedures underscore the importance of meticulous data preparation in facilitating accurate and insightful analyses.

Argument ID	Conclusion	Stance	Premise	text	category	Self-direction: thought	...	Universalism: objectivity
A26016	We should ban naturopathy	in favor of	naturopathy is very ...	naturopathy is very dangerous for the most vulnerable people, like children and cancer patients. people use ineffective treatments and forgo proven cures, such as antibiotics or chemo, often resulting in death. in favor of We should ban naturopathy	['Achievement', 'Security: personal', 'Benevolence: dependability', 'Universalism: concern']	0	...	0
A26080	We should subsidize stay-at-home dads	against	The suggestion that men ...	The suggestion that men deserve payment for what has traditionally been women's unpaid work is problematic; it sends the message that only men's labor has value. against We should subsidize stay-at-home dads	['Face', 'Tradition', 'Benevolence: caring', 'Universalism: concern', 'Universalism: objectivity']	0	...	1
A27056	We should abandon marriage	against	marriage is a very important...	marriage is a very important institution that must be kept as a bond between people that cannot be easily broken on a whim against We should abandon marriage	['Security: personal', 'Tradition']	0	...	0
A27117	We should fight urbanization	in favor of	urbanization results in ...	urbanization results in resources becoming more scarce in favor of We should fight urbanization	['Power: resources', 'Security: personal', 'Universalism: nature']	0	...	0

2.1 Data preparation for the SimpleTransformers

In preparation for training our models using SimpleTransformers, meticulous data organization is key. SimpleTransformers operates on a straightforward structure, necessitating a clear arrangement of our dataset. It's imperative to ensure that our dataset aligns with SimpleTransformers' requirements: a designated text column and a corresponding label column.
If our dataset comes equipped with a header row, it's our responsibility to ensure that our columns are appropriately labeled. The column containing our textual data should be distinctly labeled as 'text', signifying its role in housing the text inputs. Conversely, the column hous9ing our labels should be aptly labeled as 'labels', highlighting its purpose in storing the associated labels for our textual inputs.

	text	labels
0	we should ban human cloning as it will only ca...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
1	fast food should be banned because it is reall...	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...
2	sometimes economic sanctions are the only thin...	[0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
3	capital punishment is sometimes the only optio...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, ...
4	factory farming allows for the production of c...	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, ...
...	...	...
5308	On the one hand, we have Russia killing countl...	[0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
5309	The subsidies were originally intended to ensu...	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, ...
5310	These products come mainly from large enterpri...	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
5311	Subsidies often make farmers in recipient coun...	[0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, ...
5312	The EU cannot endlessly lean on America or NAT...	[0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...

3 Methods

In the methods section, we utilize three prominent language models: BERT, ALBERT, and RoBERTa. These models have become foundational components in natural language processing tasks and are widely recognized for their effectiveness in various NLP applications.

3.1 Simple Transformers

Simple Transformers is a natural language processing (NLP) library designed to streamline the use of powerful Transformer models, like BERT, RoBERTa, and ALBERT. It allows you to quickly train and evaluate these models for various NLP tasks without getting bogged down in the complexities of their underlying architectures.
In essence, Simple Transformers acts as a bridge between powerful Transformer models and everyday NLP tasks. It empowers you to harness the capabilities of these models without needing extensive deep learning knowledge.

3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a language model introduced by researchers at Google in October 2018. It is based on the transformer architecture and has significantly improved upon previous state-of-the-art models in natural language processing (NLP).
Here are some key points about BERT:

Purpose:

BERT is designed to handle various NLP tasks and serves as a versatile solution for tasks like sentiment analysis, named entity recognition, and more.

Architecture:

BERT follows an “encoder-only” transformer architecture. It consists of three main modules:
Embedding: Converts one-hot encoded tokens into vectors representing the tokens.
Stack of Encoders: These are the Transformer encoders that transform the representation vectors.
Un-embedding: Converts the final representation vectors back into one-hot encoded tokens. This module is mainly used during pretraining .

Pretraining:

BERT is pretrained on two tasks:
Masked Language Modeling: Predicting a selected token given its context. For example, given the sentence “my dog is cute,” BERT predicts the missing word (e.g.,

"[MASK]") based on context.
Next Sentence Prediction: Determining if two spans of text appear sequentially in the training corpus.

Vocabulary and Tokenization:

BERT uses WordPiece tokenization to convert English words into integer codes. Its vocabulary size is 30,000. Any token not in the vocabulary is replaced by "[UNK]" for "unknown".
In summary, BERT's bidirectional approach allows it to consider context by analyzing word relationships in sentences from both directions, making it a powerful tool for various NLP tasks.

3.3 AlBERT

The AlBERT (A Lite BERT) model is a variant of the original BERT (Bidirectional Encoder Representations from Transformers) introduced by researchers at Google.
Here are the key points about AlBERT:

Purpose:

Like BERT, AlBERT is designed to handle various natural language processing (NLP) tasks. It serves as a versatile solution for tasks such as sentiment analysis, named entity recognition, and more.

Architecture:

AlBERT follows the same fundamental architecture as BERT, which is the transformer architecture. However, AlBERT introduces a crucial modification: it uses a parameter-sharing strategy to reduce the model's size and improve efficiency. Specifically, AlBERT shares parameters across different layers, resulting in a more compact representation.

Pretraining:

AlBERT undergoes pretraining, similar to BERT. The pretraining tasks include:
Masked Language Modeling (MLM): Predicting a masked token given its context (similar to BERT's language modeling task).
Sentence Order Prediction (SOP): Determining if two spans of text appear sequentially in the training corpus (similar to BERT's next sentence prediction task).

Vocabulary and Tokenization:

AlBERT employs the same WordPiece tokenization as BERT to convert English words into integer codes. Its vocabulary size remains at 30,000 tokens. Any token not present

in the vocabulary is replaced by “[UNK]” to indicate an unknown token.
In summary, AlBERT inherits BERT’s bidirectional approach, allowing it to consider context by analyzing word relationships in sentences from both directions. This makes AlBERT a powerful tool for various NLP tasks, while its parameter-sharing strategy enhances efficiency and reduces model size.

3.4 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) builds upon BERT’s foundation. Like BERT, RoBERTa is a transformer-based language model that uses self-attention to process input sequences and generate contextualized word representations. It was proposed by researchers at Facebook AI.
Key Differences and Improvements:

Training Strategy:

RoBERTa trains for more epochs, with larger mini-batches and on a larger training corpus. It removes the next sentence prediction objective present in BERT.

Vocabulary and Tokenization:

RoBERTa uses a byte-level Byte-Pair Encoding (BPE) tokenizer (similar to GPT-2) instead of BERT’s WordPiece tokenizer.
Unlike BERT, RoBERTa doesn’t require `token_type_ids` to indicate segment boundaries, you can simply separate segments with the separation token.

Performance:

RoBERTa achieves state-of-the-art results on various benchmarks, including the GLUE, RACE, and SQuAD datasets.
In summary, RoBERTa inherits BERT’s bidirectional approach while optimizing hyperparameters and training strategies, making it a powerful tool for NLP tasks.

3.5 Evaluation metrics

The primary metric of interest is the F1 score, a widely used measure that balances a model’s precision and recall. Additionally, we consider supplementary metrics such as confusion matrices and other elements including true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Definitions:

Recall (R): Recall, also known as sensitivity, is the ratio of true positive predictions to the total number of actual positive instances. It measures the model’s ability to correctly identify positive instances from the total pool of actual positives.

Formula: R = TP / (TP + FN)

Precision (P): Precision is the ratio of true positive predictions to the total number of positive predictions. It measures the model's ability to accurately identify positive predictions, avoiding false positives.
Formula: P = TP / (TP + FP)

F1 score(F1): The F1 score is a metric that combines precision and recall into a single value, providing a balanced measure of a classification model's accuracy. It indicates how well the model can correctly identify relevant instances while minimizing false positives and false negatives.
Formula: F1 = 2 * (precision * recall) / (precision + recall)

FPR (False Positive Rate): FPR is the ratio of false positive predictions to the total number of actual negative instances. It measures the proportion of negative instances that are incorrectly classified as positive.
Formula: FPR = FP / (FP + TN)

MPR (Misclassification Precision Ratio): MPR is the ratio of misclassification errors to the total number of positive predictions. It provides an indication of the balance between false positives and false negatives in the model's positive predictions.
Formula: MPR = (FP + FN) / (TP + FP)

MBR (Misclassification Bias Ratio): MBR is the ratio of misclassification errors to the total number of instances. It provides an indication of the balance between false positives and false negatives in the model's predictions. A lower MBR suggests a more balanced performance.
Formula: MBR = min(FP / (FP + TN), FN / (FN + TP))

4 Result and discussion

We trained three different models for our task: BERT, ALBERT, and RoBERTa. Each model was evaluated using the F1 score on four different datasets, a widely used metric that balances a model's precision and recall. Below are the results of our experiments:

BERT Results and F1 Scores Across any Database

After training and evaluating the BERT model on our datasets, we obtained the following results:

	Unnamed: 0	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	...
0	First_Data	0.407672	0.492537	0.567527	0.049383	0.206897	0.553020	0.251656	0.492308	0.058252	...
1	Zhihu_Data	0.370299	0.500000	0.432432	1.000000	0.000000	0.625000	0.000000	0.478261	0.000000	...
2	NHJ_Data	0.177066	0.040000	0.140845	0.000000	0.666667	0.459016	0.000000	0.285714	0.000000	...
3	NYT_Data	0.312219	0.571429	0.000000	1.000000	0.000000	0.428571	0.000000	1.000000	0.000000	...

F1-Score Values Across all the Datasets with BERT model.

AlBERT Results and F1 Scores Across any Database.

After training and evaluating the ALBERT model on our datasets, we obtained the following results:

	Unnamed: 0	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	...
0	First_Data	0.363518	0.403042	0.535662	0.047619	0.125	0.473920	0.223684	0.359551	0.108108	...
1	Zhihu_Data	0.349825	0.133333	0.400000	1.000000	0.000	0.621622	0.000000	0.382979	1.000000	...
2	NHJ_Data	0.183367	0.071429	0.160000	0.000000	0.400	0.460317	0.000000	0.400000	0.100000	...
3	NYT_Data	0.241290	0.000000	0.000000	1.000000	0.000	0.258065	0.285714	1.000000	0.000000	...

F1-Score Values Across all the Datasets with AlBERT model.

RoBERTa Results and F1 Scores Across any Database.

After training and evaluating the RoBERTa model on our dataset, we obtained the following results:

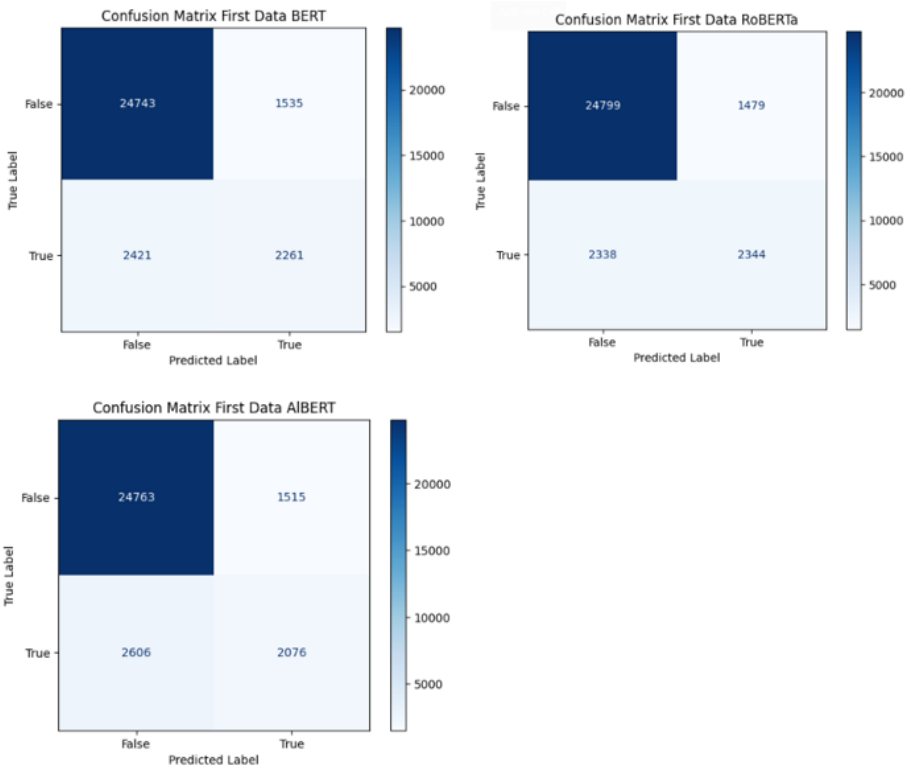
	Unnamed: 0	All	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	...
0	First_Data	0.435849	0.472868	0.608824	0.04878	0.312500	0.587450	0.315789	0.419355	0.183206	...
1	Zhihu_Data	0.409568	0.380952	0.484848	1.00000	0.666667	0.707317	0.000000	0.416667	0.500000	...
2	NHJ_Data	0.231624	0.097561	0.193548	0.00000	0.500000	0.544118	0.000000	0.000000	0.258065	...
3	NYT_Data	0.298394	0.571429	0.200000	1.00000	0.000000	0.250000	0.000000	1.000000	0.000000	...

F1-Score Values Across all the Datasets with AlBERT model.

4.1 Further Evaluation of Each Database

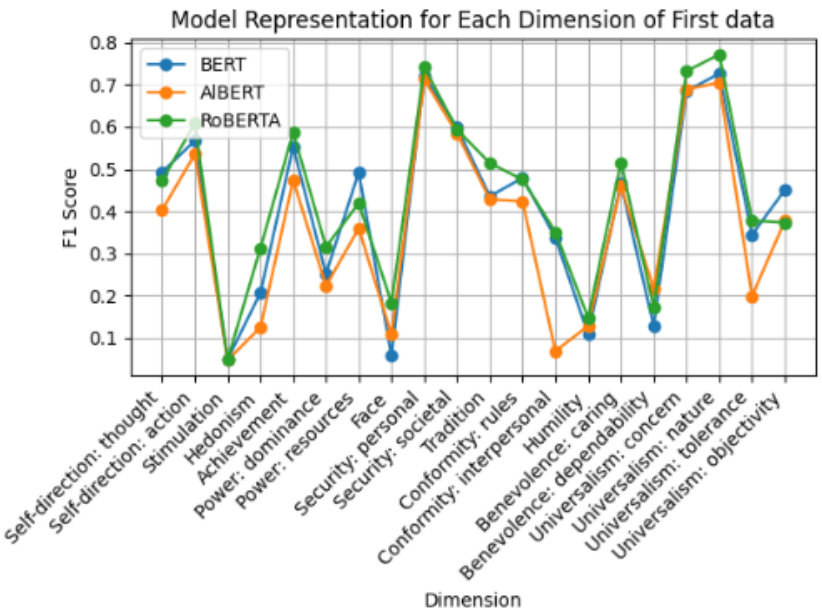
First Dataset

To visually compare the performance of each model on the First Dataset, we'll generate plots representing the confusion matrices of each model:

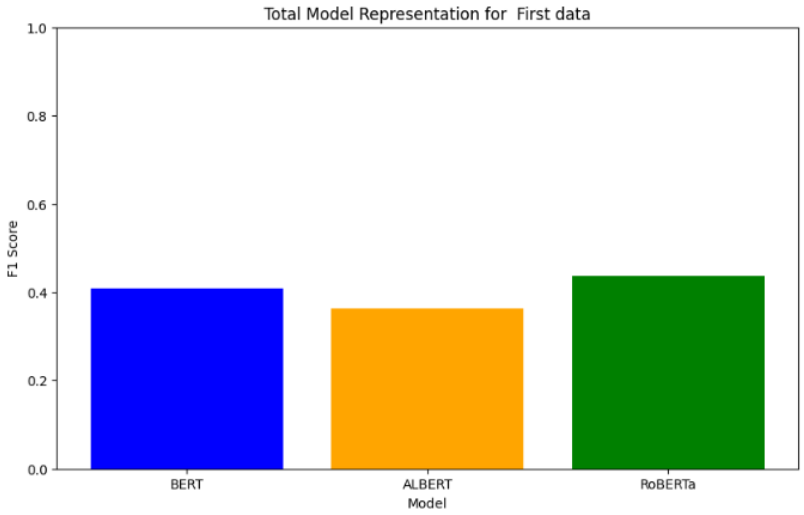


Comparison of of Confusion Matrix all models for the First dataset.

For a more comprehensive assessment of model performance, we offer the following plot. It illustrates the F1 scores of each model across all labels, aiding in comparative analysis.After the plot, you will find another visualization representing the average F1 score of all models. This provides a more thorough assessment of model performance.



Comparison of individual label F1-Score of all models for the First data

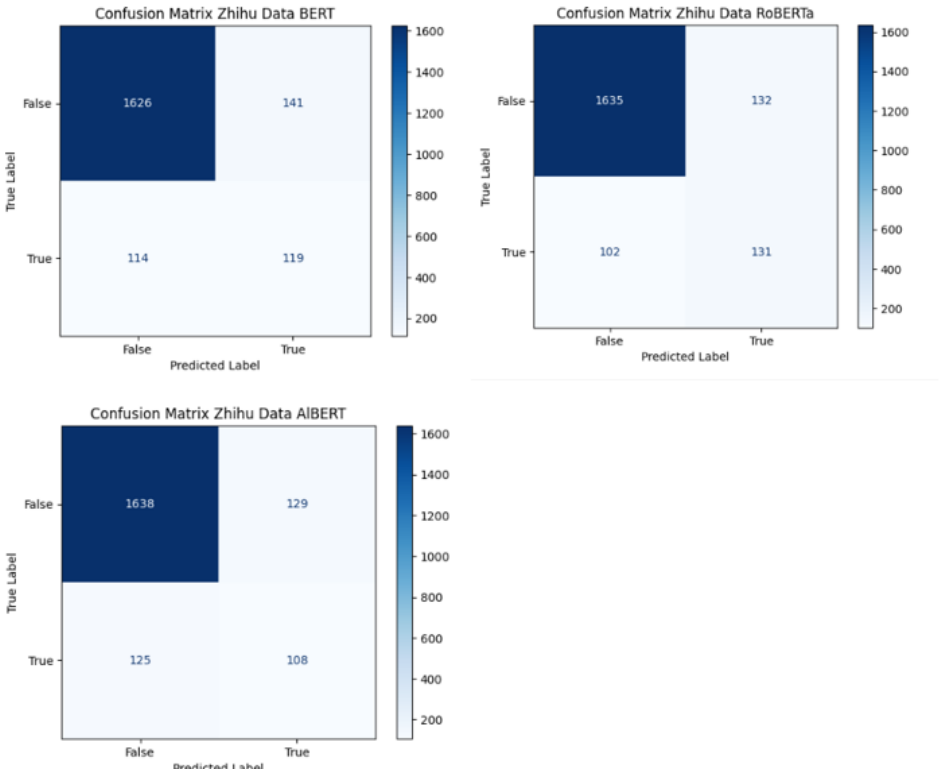


Comparison of average F1-Score of all models for the First Dataset.

This additional comparison highlights RoBERTa's dominance on the First dataset. It's evident from both the overall F1-score and the F1-scores for each specific label that RoBERTa consistently outperforms BERT and ALBERT.it maintains the highest recall, and precision, as well as the highest MBR and MPR.(MBR = 0.3265, MPR = 0.5252)

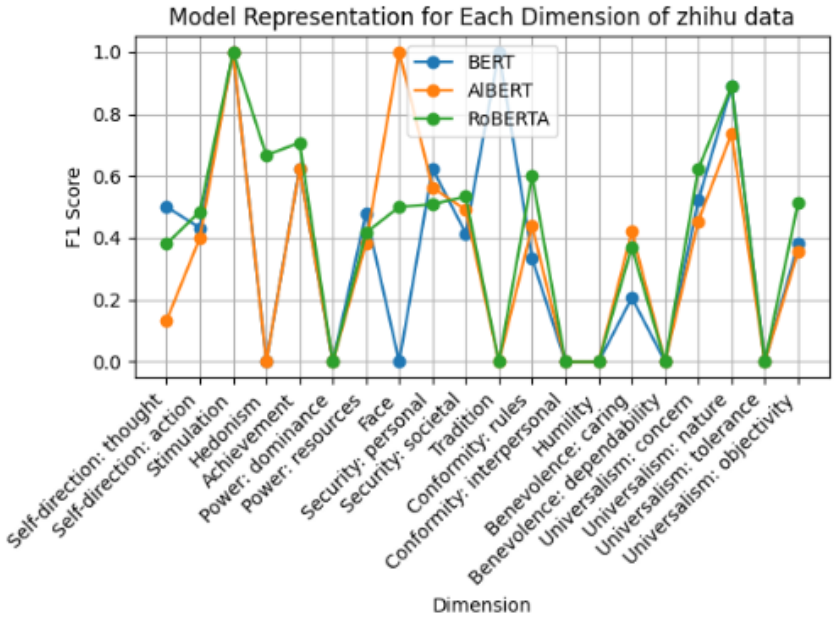
Zhihu Dataset

To effectively compare how each model performs on the Zhihu Dataset, we will generate plots that depict the confusion matrices for each model:

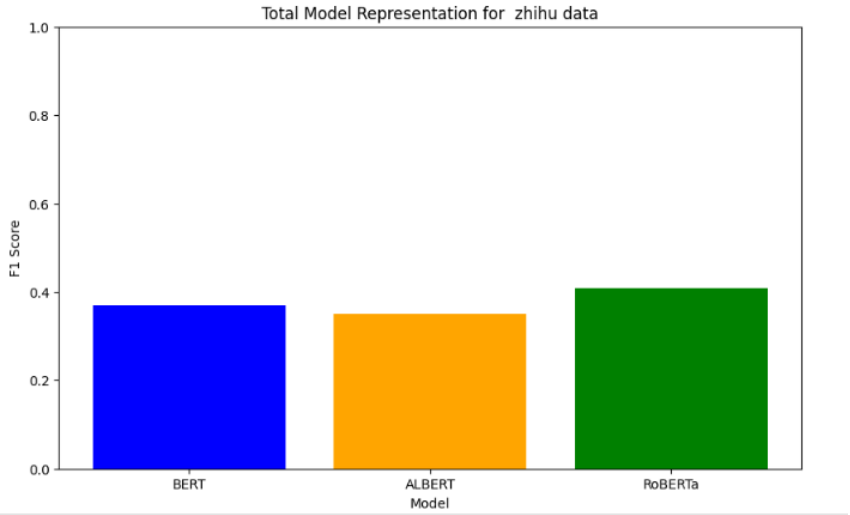


Comparison of of Confusion Matrix all models for the Zhihu dataset.

The plot illustrates F1 scores for each model across labels. Afterward, there's another visualization showing the average F1 score for all models.



Comparison of individual label F1-Score of all models for the Zhihu data.

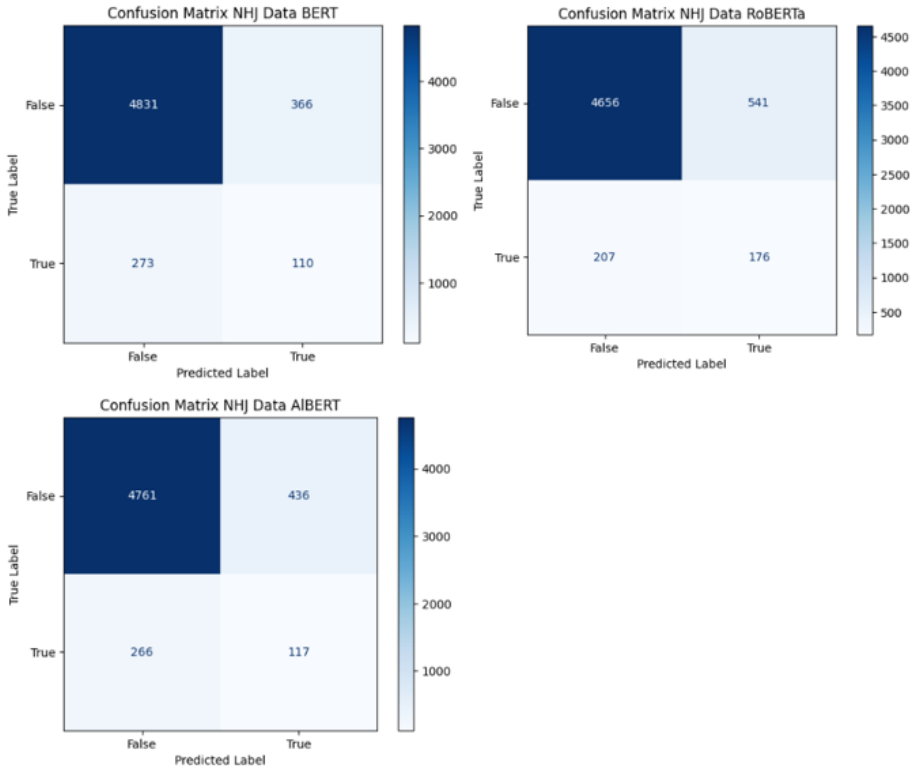


Comparison of average F1-Score of all models for the the Zhihu Dataset.

RoBERTa again performs the best across all metrics, with the highest F1 score, recall, and precision and it also remains the top performer with the highest MBR and MPR.(MBR = 0.2939, MPR = 0.5337)

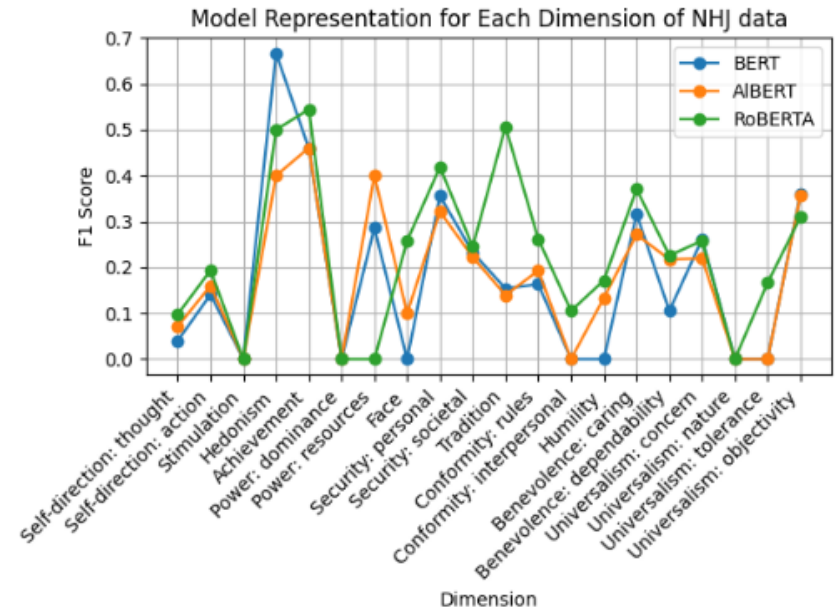
Nahjalbalagh Dataset

To facilitate a visual comparison of the performance of each model on the new dataset, we will generate plots that depict:

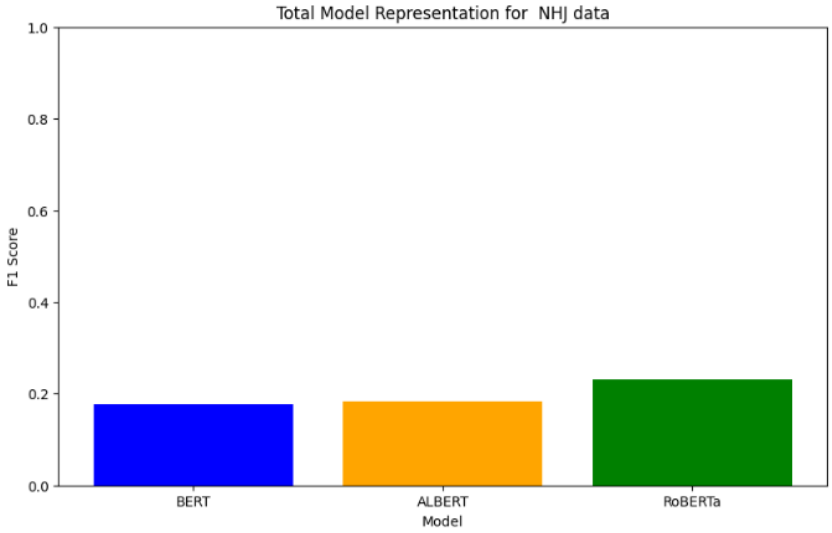


Comparison of of Confusion Matrix all models for the Nahjalbalagh dataset.

The plot illustrates F1 scores for each model across labels. Afterward, there's another visualization showing the average F1 score for all models.



Comparison of individual label F1-Score of all models for the Nahjalbalagh dataset.

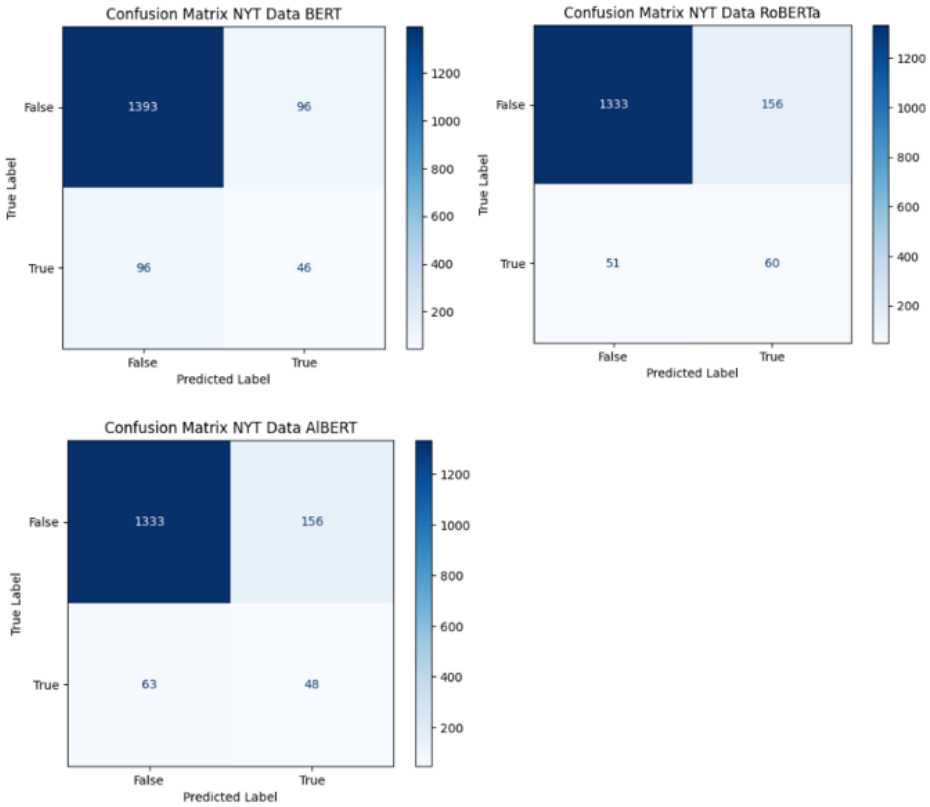


Comparison of average F1-Score of all models for the the Nahjalbalagh data.

RoBERTa once again outperforms the other models in terms of F1 score, recall, precision, MBR and MPR.(MBR = 0.1560, MPR = 0.3439)

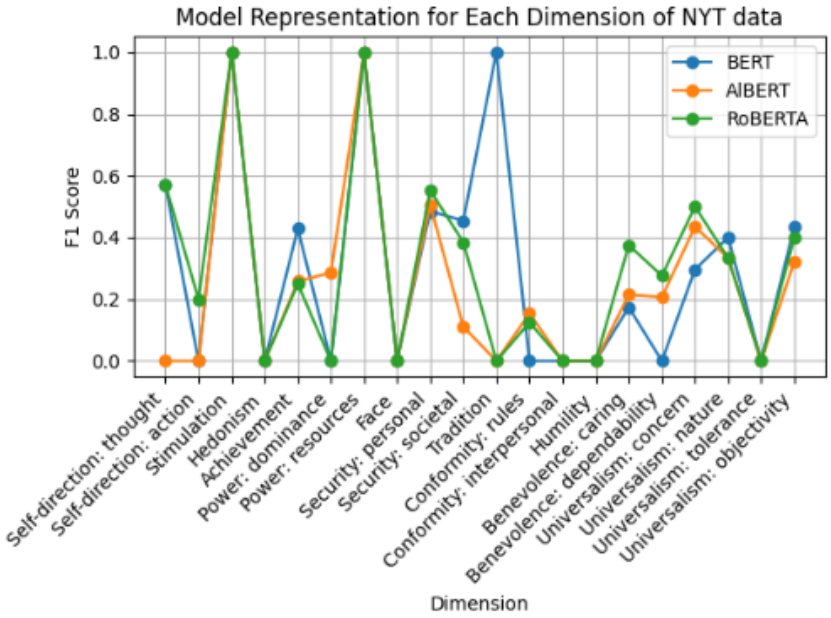
New York Times Dataset

To simplify the comparison of model effectiveness on the New York Times dataset, we will craft visual aids, revealing insights into:

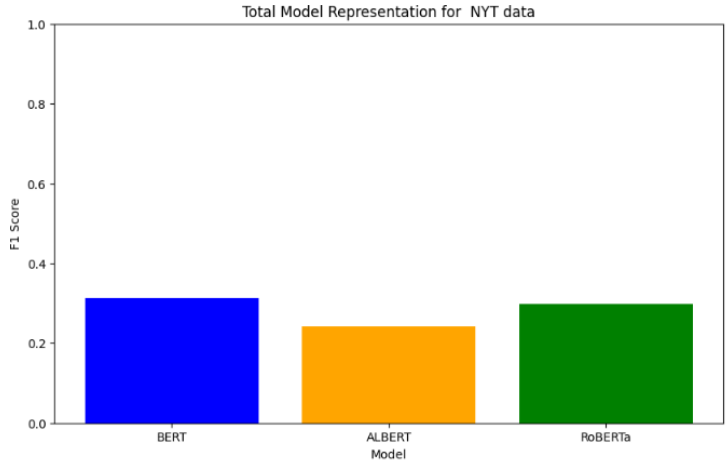


Comparison of of Confusion Matrix all models for the New York Times dataset.

The plot illustrates F1 scores for each model across labels. Afterward, there's another visualization showing the average F1 score for all models.



Comparison of individual label F1-Score of all models for the New York Times dataset.



Comparison of overall F1-Score of all models for the New York Times dataset.

In the NYT Data dataset, BERT stands out as the top-performing model. Despite RoBERTa's competitive performance, BERT achieves the highest precision value among all models, showcasing its accuracy in identifying relevant instances. Additionally, BERT maintains a commendable F1 score, indicating a good balance between precision and recall. While RoBERTa demonstrates strong overall performance, BERT's superior precision and competitive F1 score make it the preferred choice for handling the NYT Data dataset.
Furthermore, when comparing individual label F1-Scores for the New York Times dataset, interesting patterns emerge. RoBERTa exhibits a more consistent behavior across different labels, whereas BERT shows fluctuations. However, despite this variability, BERT's exceptional precision and competitive F1 score still position it as the optimal model for processing the NYT Data dataset.

In summary, RoBERTa consistently displays robust performance across various datasets.

4.2 Summery of Results

Dataset	Model	Precision (P)	Recall (R)	F1 Score (F)
First Data	BERT	0.5496	0.3237	0.4077
	ALBERT	0.4775	0.2940	0.3635
	RoBERTa	0.5485	0.3618	0.4358
Zhihu Data	BERT	0.5513	0.2785	0.3703
	ALBERT	0.5489	0.2567	0.3498
	RoBERTa	0.5008	0.3467	0.4096
NHJ Data	BERT	0.5295	0.1063	0.1771
	ALBERT	0.2574	0.1422	0.1834
	RoBERTa	0.2449	0.2195	0.2316
NYT Data	BERT	0.6270	0.2079	0.3122
	ALBERT	0.3067	0.1993	0.2413
	RoBERTa	0.3488	0.2607	0.2984

In evaluating the performance of different models across the datasets, RoBERTa consistently emerges as the top performer. Its precision, recall, and F1 scores exhibit competitive values, often surpassing those of other models such as BERT and ALBERT. Therefore, based on the provided metrics, RoBERTa demonstrates superior effectiveness in handling the datasets compared to its counterparts.

5 Conclusion

In conclusion, our project encompassed the comprehensive evaluation of three prominent models BERT, ALBERT, and RoBERTa across a diverse range of datasets, including First, Zhihu, Nahjalbalagh and New York Times.

We meticulously prepared the data and trained each model, evaluating their performance using F1 scores and confusion matrices.

Across all datasets, RoBERTa consistently demonstrated superior performance compared to the other models. Its consistently high F1 scores and effective management of our task, underscore RoBERTa's suitability for our project

In conclusion, RoBERTa emerges as the top-performing model across different datasets, making it the preferred choice for our multi-label classification task due to its robust performance and adaptability.