SCaLAR NITK at Touché: Comparative Analysis of Machine Learning Models for Human Value Identification

Notebook for the Touché Lab at CLEF 2024

Praveen K[1,†], Darshan R K[2,‡], Chinta Tejdeep Reddy[3,§] and Anand Kumar M[4,¶]

[1]National Institute of Technology Karnataka, Surathkal, Mangaluru, India

Abstract

This study delves into task of detecting human values in textual data by making use of Natural Language Processing (NLP) techniques. With the increasing use of social media and other platforms, there is an abundance in data that is generated. Finding human values in these text data will help us to understand and analyze human behavior in a better way, because these values are the core principle that influence human behavior. Analyzing these human values will help not only in research but also for practical applications such as sentiment evaluation, market analysis and personalized recommendation systems. The study tries to evaluate the performance of different existing models along with proposing novel techniques. Models used in this study range from simple machine learning model like SVM, KNN and Random Forest algorithms for classification using embeddings obtained from BERT till transformer models like BERT and RoBERTa for text classification and Large Language Models like Mistral-7b. The task that has be performed is a multilabel, multitask classification. QLoRA quantization method is used for reducing the size of weights of the model which makes it computationally less expensive for training and Supervised Fine Tuning (SFT) trainer is used for fine tuning LLMs for this specific task. It was found that LLMs performed better compared to all other models.

Keywords

Human Values, SVM, BERT, RoBERTa, Mistral, SFT trainer, QLoRA

1. Introduction

In the present world, where everyone is digitally connected, huge volumes of text data are generated and shared across various platforms including social media, scholarly articles etc. Within each of these texts lie implicit indicators of human values which are core principles and beliefs that guide individuals' actions, perceptions and decisions as mentioned in [1]. These human values include stimulation, hedonism, achievement, dominance, humility etc. Detecting and understanding human values in text is not only essential for market research, brand sentiment analysis, and political discourse analysis but also for applications such as personalized recommendation systems, content filtering, and social media monitoring. Even though these values are often implicit, these values manifest in language through expressions, sentiments and contextual cues. However, the automation of these human value detection is a challenging task due to their abstract and subjective nature. Recent advancement in NLP have opened up new possibilities for understanding languages at deeper levels.

In recent years NLP have been extensively used for tasks such as information retrieval, sentiment analysis etc. In this paper, we focus on the task of detecting human values in text using state-of-the-art NLP techniques. We aim to explore methodologies for automating this task by identifying and categorizing values expressed in textual data. Throughout this paper, we delve into challenges associated with detecting human values by discussing exiting approaches and methodologies. We the propose novel techniques to increase accuracy and enhancing the performance of the model, with experiment results. The study includes identifying human values in eight different languages. We make use of existing language translation models to translate non-English sentences to English and then perform classification.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

† Authors contributed equally

praveenk.211ai001@nitk.edu.in (P. K); darshanrk.211ai015@nitk.edu.in (D. R. K); tejdeep.211ai013@nitk.edu.in (C. T. Reddy); m\protect1_anandkumar@nikt.edu.in (A. K. M)

2. Background

In addition to what is currently known about the identification of human values within arguments, the researchers at Johannes Kiesel and colleagues' effort "ValueEval: Identification of Human Values Behind Arguments" [2] review the literature. They look into past studies and already defined approaches in this area to provide a foundation for their own research. In an attempt to understand how human values will be expressed and meaning will be derived in discussions, they go through a list of sources. Through this literature study, they intend to close these knowledge and methodological gaps by developing their own ValueEval methods. In order to group different perspectives on the topic, the researchers find a range of disciplines, including psychology, linguistics, and computer science. Through this depth focused research, they are trying to provide readers a greater understanding of the role that human values play in arguments as well as how to identify and evaluate them. By combining the body of previous research, the researchers expect to advance the development of techniques for understanding and evaluating the values.

Machine learning research frequently promotes performance more additional important factors like cost and energy use. Using the LexGLUE benchmark, the study compares older versions (like SVM) with more recent models (like BERT and GPT2)[3]. During the testing and manufacturing facilities phases, they evaluate timing, energy consumption, costs, and performance. Simpler models can frequently outperform more intricate ones while consuming less effort and resources. This implies that while selecting machine learning solutions, businesses should consider these extra considerations. It also weights on how crucial it is to take energy use into account when evaluating models in order to provide findings that are relevant.

M Garg et al. [4] shows the rich concept of well-being as it is presented in Reddit discussion threads, which covers elements like mental and social well-being. The MultiWD dataset has 3281 annotated rows and that they carefully selected for the purpose for them to understand and find these multiple dimensions. This dataset is a useful tool for looking at wellness indicators of online content. Using advanced classifiers, they carefully saw a number of models; the optimized BERT models excelled the others, with an F1 score of 76.69. This shows how easy it is to find complex features of wellness in created by users writing with modern techniques, such as optimizing language models. Their results point out the significance it is to use domain-specific data to enhance AI.

M. Granitzer et al. [5] explained the hierarchical structure of text classification is represented by a directed acyclic graph, and the application of Boosting in this context is examined. It contrasts the performance of Boosting methods with Support Vector Machines (SVMs). As a hierarchical grouping mechanism works its way down from the top, each node decides whether or not to distribute a document farther. Flat classifiers like SVMs, CentroidBooster, and BoosTexter are used at each node. CentroidBooster is an AdaBoost.MH-based BoosTexter solution. An examination of the Reuters Corpus Volume 1 and the OHSUMED datasets shows that recognizing the hierarchical structure of a dataset increases the F1-measure.

F. Rollo et al. [6] showed how they faced the difficulty task of classifying Italian newspapers which becomes even harder by the language's unique structure and style. Their methodology consists of a number of processes including preprocessing the text creating word embeddings feature engineering and document vector training. They test the model's performance on a separate data-set to determine how nicely it handles fresh information. They compare eight models, look into fifteen classifiers and analyze 3 word embedding techniques. Their research includes 6 new Italian models trained on native datasets in addition to popular models like Word2Vec and FastText. The usage of an Italian GloVe model which was previously unavailable is a significant addition. They test using datasets containing news articles regarding crimes and general Italian news.The findings shows the negatives of the Decision Tree, Bernoulli, and Gaussian Naive Bayes models as well as the efficiancy of the Support Vector Classification algorithm. When word embedding models are compared, Word2Vec and GloVe perform better than FastText. All things are considered and their work improved the text classification for Italian texts and added new Italian word embedding models to the language environment.

In recent times people have shown worriness about the mean and hurtful things said on social media

especially when it comes to women. Researchers are trying to find ways to make social media safer for everyone. The SemEval-2023 [7] Task 10 focused on finding and explaining these hurtful comments, known as sexism, on sites like Gab and Reddit. They wanted to understand the different types of sexism people face online. To do this they used special Large Language models called like XLM-T and HateBERT which were trained to recognize offensive language. These models are trained on lots of tweets and Reddit posts so they are very good at understanding how people message each other online. Kirk and their team understood and explained how they used these models to find and classify the sexist comments. The hope that by understanding sexism better they can make the internet a nicer place for everyone.

3. System Overview

3.1. Machine Learning Models (KNN, SVM, Decision Trees, Hierarchical Classification):

SVMs: SVMs are a powerful supervised learning algorithms which are mainly used for classification task [8]. It mainly works on the basis of creating hyperplane that best separates classes in the feature space. The margin maximization is used to maximize the distance between planes keeping the scores as high as possible at the same time.

KNN: KNN is a simple yet effective non-parametric algorithm used for classification and regression tasks [9]. K-Nearest Neighbors (KNN) functions by assigning the majority class of the K nearest neighbors to the query point in the feature space.

Decision Trees: These ML models are widely use for regression and classification tasks and they operate in a recursive manner [10]. Each of these models recursively splits the feature space based on feature values, selecting the feature that provides the highest information gain at each node for splitting.

SVMs, KNN and decisions trees are considered as base models and we have Used One Vs Rest classifier, as it creates a binary classifier for each class. This approach allows us to treat each class separately and provides more precise weight-age to each one. Each of the classifier learns to predict 1 or 0 for each of the target class i.e 38 classes that are present in the dataset.

Hierarchical Classification: In this model, we will classify data into three main hierarchical levels, resulting in three columns or features. Each column will have multiple values, as this is a multi-class, multi-label classification task. The first level of classification consists of four labels: Self-Transcendence, Openness to Change, Self-Enhancement, and Conservation. Each first-level label is further divided into sub-level labels. Under Self-Transcendence, we have Benevolence Caring, Benevolence Concern, Benevolence Dependability, Universalism Concern, Universalism Nature, and Universalism Tolerance. Under Openness to Change, we have Self-Direction Thought, Self-Direction Action, Stimulation, and Hedonism. Under Self-Enhancement, we have Achievement, Dominance, Resource-Power, and Face. Under Conservation, we have Humility, Conformity Interpersonal and Rules, Tradition, Security Societal, and Security Personal. Each sub-level node is further divided into "Attained" or "Constrained" as the third level. We will use one classifier per node at each level for classification. For each sentence, the labels will represent the path from the parent class to the leaf nodes. There can be multiple paths for each sentence, reflecting the multi-label nature of the classification task.

Figure 1 shows the pictorial representation of the tree structure built for hierarchical text classification model. Note that all the classes are not included in the image but are considered in the model.

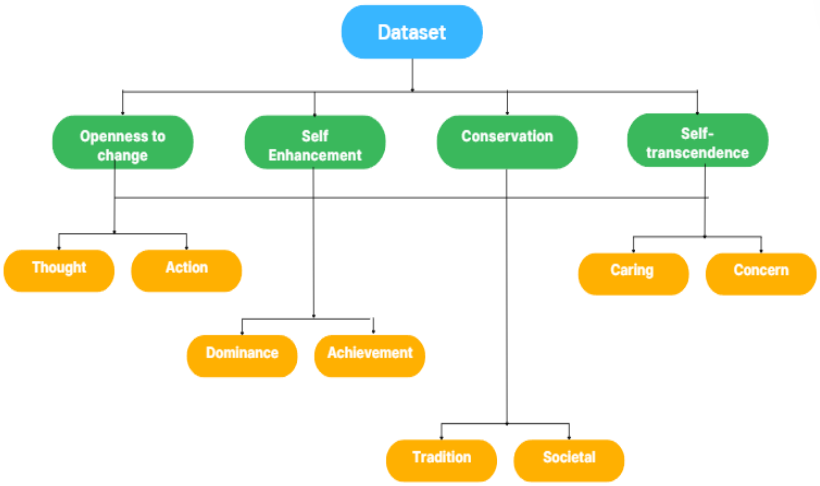

Figure 1: *Pictorial representation of the tree tructure built for hierarchical text classification*

3.2. Transformer Models (BERT,RoBerta) :

BERT [11] is fine-tuned for Multi-Label Classification. Here, only the Pooler Layer is unfrozen, while all the other layers remain frozen. There is only hidden layer and one output layer neural network architecture that is connected to the BERT embedding layer. Tokenized inputs are given to the model for training, training is done with AdamW optimizer and BCEWithLogitsLoss criterion. The training process is run for 5 epochs.
Roberta [12] is the final model which we took into considerations as it gave the highest results. The first layer was the Roberta Embedding layer. Then a dropout layer with 0.3 probability was added. Then a final layer for 38 classes was added. We have used 5 epochs with a batch size 16 and learning rate of 2e-5. We had a custom loss function which combines Cross entropy loss with focal loss components to handle class imbalance and hard examples. We have used 3 parameters alpha = 0.25 , beta = 0.5 and gamma = 2 to weight positive and negative examples.

3.3. Large Language Models (Mistral):

Mistral-7B model have been used in this study for human value detection. Quantized versions of these models are used in order to cope up with computational costs. Prompts for each text is generated before feeding it into LLMs. Since we are using SFT trainer the prompts along with class labels are fed into the model.

Figure 2 represents the flowchart indicating the flow of PEFT technique. Parameter Efficient Fine Tuning (PEFT) parameter such as regularization parameter, dropout rate and task type set to CASUAL_LM (Casual Language Modelling) are defined in a function that is used for this purpose. Various training parameters such as learning rate, weight decay rate, warmup ratio, gradient accumulation steps etc have been predefined. Adam 32-bit optimizer has been used while training. Supervised Fine Tuning (SFT) trainer is used for training because of its low memory training capability for LLMs which uses Low Rank Adaptive (LoRA) configuration for model parameters, enabling targeted parameter updates by reducing the memory required.

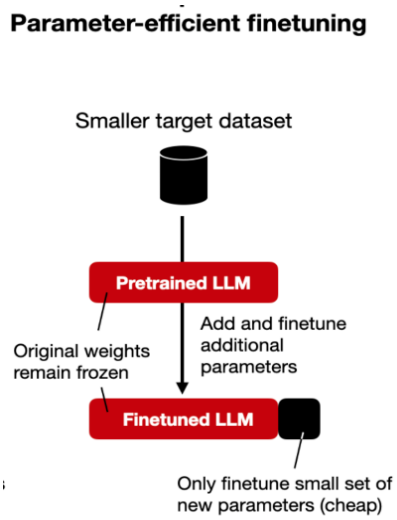

Figure 2: *Parameter Efficient Fine-tuning (PEFT)*

4. Results

All the metrics are the weighted average values for the Validation Set.

Table 1
Performance metrics for task1

	Precision	Recall	F1-score
KNN	0.08	0.07	0.07
SVM	0.11	0.10	0.10
Decision trees	0.10	0.07	0.08
BERT	0.13	0.14	0.13
Hierarchial	0.17	0.23	0.19
RoBERTa	**0.20**	**0.36**	**0.26**
Mistral	0.27	0.10	0.16

Table 2
Performance metrics for task2

	Precision	Recall	F1-score
KNN	0.03	0.05	0.04
SVM	0.07	0.10	0.08
Decision trees	0.06	0.09	0.07
BERT	0.09	0.28	0.13
Hierarchial	0.14	0.18	0.16
RoBERTa	**0.18**	**0.37**	**0.24**
Mistral	0.21	0.09	0.13

We can observe from the above table RoBERTa is the best performing models followed by hierarchical models for both the tasks. The machine learning models are found to perform poorly because they are not able to generalize better because the dataset size is of (40000 x 768) because we are using BERT embeddings which are then fed into ML models. We also observe that the scores are comparatively low because of the datset being in eight different languages and due to the absence of single model which

Table 3
Achieved F_1-score of each submission on the test dataset for subtask 1. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

		F_1-score																		
Submission	**EN**	**All**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
electric-basket-2024-05-06-16-22-15	✓	28	05	17	27	27	38	34	38	15	34	40	41	43	07	00	23	26	37	56	16
valueeval24-bert-baseline-en	✓	24	00	13	24	16	32	27	35	08	24	40	46	42	00	00	18	22	37	55	02
valueeval24-random-baseline		06	02	07	05	02	11	08	10	04	05	13	03	11	03	00	04	04	09	04	02

Table 4
Achieved F_1-score of each submission on the test dataset for subtask 2. A ✓ indicates that the submission used the automatic translation to English. Baseline submissions shown in gray.

		F_1-score																		
Submission	**EN**	**All**	**Self-direction: thought**	**Self-direction: action**	**Stimulation**	**Hedonism**	**Achievement**	**Power: dominance**	**Power: resources**	**Face**	**Security: personal**	**Security: societal**	**Tradition**	**Conformity: rules**	**Conformity: interpersonal**	**Humility**	**Benevolence: caring**	**Benevolence: dependability**	**Universalism: concern**	**Universalism: nature**	**Universalism: tolerance**
valueeval24-bert-baseline-en	✓	81	83	79	86	88	84	77	80	74	84	81	78	78	79	87	89	86	85	81	78
electric-basket-2024-05-06-16-22-15	✓	77	69	72	78	73	79	77	79	71	78	81	79	77	70	70	77	76	79	80	71
valueeval24-random-baseline		53	55	49	52	54	52	56	56	50	48	54	50	54	55	61	55	51	48	51	51

can perform well on all languages.

When the same task is performed on english dataset alone, the metrics are found to be better than the above. The high metric was achieved with the help of Mistral-7b which gave precision 0.52, recall of 0.20 and f1-score of 0.28. his suggests that the model is able to predict the positive classes properly, in this task it is predicting classes which are actually true, but it is also predicting classes which are not supposed to be present. Roberta has given the highest score when validated only english dataset displaying weighted metrics as 0.28 for precision, 0.47 for Recall and 0.35 for f1-score.

Table 3 represents the results of our approach compared to the base model for subtask 1, which indicates that it performs better than baseline models. Table 4 represents the results of our approach for subtask 2, which indicates that it performs better than 2 baseline models but not as good as bert baseline.

5. Conclusion

We initially started with conventional ML models like SVMs, KNNs and decision Trees with Multilingual BERT embeddings for texts. We observed low metrics because of the existence of multiple languages and the model wasn't able to generalize. The same Conventional ML models showed a jump of 8-10% in all metrics considering just English and UnCased-BERT for embeddings. The same thing is observed

with Mistral-7b fine-tuning and also Hierechial Model. Finally it was observed that RoBERTa was the best performing model compared to other models. We aim to develop separate classification model for each of the languages. In this way we can try to increase the efficiency of the final model, because it was observed that the other language texts were pulling down the accuracy of the model which was developed. We can also create an ensemble model where we can have multiple LLMs and the final output is consider based on maximum voting which further decreases the bias created by a single model, thereby increasing its accuracy.

References

[1] J. Kiesel, Ç. Çöltekin, M. Heinrich, M. Fröbe, M. Alshomary, B. D. Longueville, T. Erjavec, N. Handke, M. Kopp, N. Ljubešić, K. Meden, N. Mirzakhmedova, V. Morkevičius, T. Reitis-Münstermann, M. Scharfbillig, N. Stefanovitch, H. Wachsmuth, M. Potthast, B. Stein, Overview of Touché 2024: Argumentation Systems, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[2] N. Mirzakhmedova, J. Kiesel, M. Alshomary, M. Heinrich, N. Handke, X. Cai, V. Barriere, D. Dastgheib, O. Ghahroodi, M. Sadraei, E. Asgari, L. Kawaletz, H. Wachsmuth, B. Stein, The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), International Committee on Computational Linguistics, 2024.

[3] L. Rigutini, A. Globo, M. Stefanelli, A. Zugarini, S. Gultekin, M. Ernandes, Performance, energy consumption and costs: a comparative analysis of automatic text classification approaches in the legal domain, INTERNATIONAL JOURNAL ON NATURAL LANGUAGE COMPUTING, 13(1), 19-35. (2024).

[4] M. Garg, X. Liu, M. Sathvik, S. Raza, S. Sohn, Multiwd: Multi-label wellness dimensions in social media posts, Journal of Biomedical Informatics 150 (2024) 104586.

[5] M. Granitzer, P. Auer, Experiments with hierarchical text classification, in: Proc. of 9th IASTED International Conference on Artifical Intelligence, 2005.

[6] F. Rollo, G. Bonisoli, L. Po, A comparative analysis of word embeddings techniques for italian news categorization, IEEE Access 12 (2024) 25536–25552.

[7] S. M. Aliyu, I. Abdulmumin, S. H. Muhammad, I. S. Ahmad, S. A. Salahudeen, A. Yusuf, F. I. Lawan, Hausanlp at semeval-2023 task 10: Transfer learning, synthetic data and side-information for multi-level sexism classification, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 1983–1987.

[8] M. Hearst, S. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intelligent Systems and their Applications (1998).

[9] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory (1967).

[10] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees, IEEE Transactions on Pattern Analysis and Machine Intelligence (1984).

[11] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding (2019).

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach (2019).