Exploring Automatic Translation between Affect Representation Schemes
Text Content Analysis

Mira Ilieva

Supervisor(s): Bernd Dudzik, Chirag Raman

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Mira Ilieva
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Chirag Raman, Alan Hanjalic

Abstract

This research delves into the exploration of translation methods between affect representation schemes within the domain of text content analysis. We assess their performance on various affect analysis tasks while concurrently developing a robust evaluation framework. Furthermore, we collect annotated datasets and take into account crucial contextual and individual factors. Ultimately, our goal is to contribute to the advancement of powerful and sophisticated tools for affect analysis. We believe a successful automated translation will aid in achieving a more comprehensive and rounded understanding of affect and further research in different fields, such as psychology and sociology.

1 Introduction

Affective content analysis is the process of identifying and extracting emotions, moods, and sentiments expressed in natural language text. It has a wide range of applications in a variety of different areas, however, the representation of affective content is still a challenging task, as emotions are often subjective and difficult to showcase in a singular representation. There is a number of different representation schemes which have been used to represent emotions in a systematic manner [1].

In recent years, there has been a surge of interest in extracting sentiment from a diverse range of textual sources, spanning from tweets to published literature [2]. These findings have opened up new possibilities and are being leveraged in various fields, including politics, finance, and education [3]. The exploration of sentiment analysis has proven to be a valuable asset, empowering advancements and generating impactful insights across multiple domains. However, this information is often represented in a singular representation scheme [4] [5], which gives a limited insight of the text, and is sometimes not readable to humans, which limits its reach. In those cases, a translation agent would add to the comprehension of the text and add robustness to the extracted sentiments, as well as enhance the overall effectiveness and applicability of sentiment analysis in practical settings.

Related Work

Furthermore recent works have been published in the field of mappings between representation schemes. A prime example is the Landowska et al research [6] where linear regression is used to map categorical emotions to dimensional (valence, arousal, dominance) using a linear regression technique. The study discusses mapping accuracy evaluation and proposes new mapping matrices for emotion representation models, however, it does not specifically address the topic of automatic translation between affect representation schemes. Similarly, in the works of Hinojosa et al [7] and Kapucu et al [8] the correlation between valence and arousal, and the categorical model is explored.

Additionally, on the topic of gender related generalisability, there is already some work done by Bauer et al [9] which explores how men and women tend do have different overall ratings across distinct categories. However, it should be noted that there is currently no comprehensive and universally accepted framework for automatic translation between affect representation schemes.

Problem Statement

The aim of this research is to investigate the availability of affective datasets annotated with multiple affect representations, evaluate the performance of various machine learning models for translating between these representations. We will examine the influence of relevant dataset properties on the generalisation capacity of translation models, including factors such as the cultural disparities arising from the construction of the dataset in different countries. Additionally, we will assess feasibility of cross gender translations feasibility and the effect of gender on the accuracy of automatic translation between affect representation schemes.

Gaining a comprehensive understanding of these aspects is pivotal in addressing the challenges associated with affect representation schemes translation, as well as uncovering potential biases or limitations within existing models. By exploring these questions, this research seeks to contribute to the field of affective computing and enhance the development of accurate and unbiased affect representation translation systems.

The findings from this study will shed light on the state-of-the-art affective datasets, evaluation procedures, relevant dataset properties, optimal machine learning approaches, and potential gender-related variations in affect representation translation [10]. This knowledge will assist researchers and practitioners in making informed decisions regarding the affect representation choice, model performance evaluation, mitigating gender biases in affect-related applications. Furthermore, this research will highlight potential limitations associated with the task, providing a comprehensive understanding of the field and facilitating further advancements.

Contribution

In this paper, we perform a comprehensive comparison between the Majority classifier and other classification models to accurately estimate the translation between affect representation schemes.

Namely, they are, from the scikit[1] library - Logistic Regression, Decision Tree, Random Forest and ANN (MLPClassifier). The Majority Classifier serves as a benchmark to evaluate the classification performance of the other models. The primary objective of this study is to thoroughly explore the feasibility of translating affect representation schemes, while also taking potential bias considerations into account. Additionally, we aim to assess the effectiveness of different classification models.

To achieve these objectives, we examine the performance of the machine learning models on three different scenarios: a single dataset (trained and tested on the same dataset), cross-dataset evaluation, and cross-gender evaluation. This evaluation allows us to gain insights into the robustness and gener-

[1] https://scikit-learn.org

alisation capacity of the classification models for affect representation translation.

Structure

The structure of this paper is as follows: Section 2 provides a detailed explanation of the methodological approach, including the datasets used, labels assigned, and the machine learning models explored. In Section 3, we outline the experimental setup and describe how the translation process was conducted. Section 4 presents the results obtained from the experiments, focusing on each individual machine learning model. Section 5 is dedicated to addressing any edge cases and offering additional insights on the topic. Section 6 summarises the findings and presents a concluding remark. Ethical implications related to the use of real-life data are discussed in Section 7. Lastly, Section 8 offers suggestions for future improvements and provides recommendations for further research.

2 Methodology

This section aims to provide information on the approach undertaken to conduct our experiment. Specifically, it delves into the used *datasets*, particular *representation schemes* (labels), *dataset properties* and data *pre-processing machine learning models* explored in our study.

2.1 Datasets selection

The nature of the task at hand necessitates the utilisation of a textual dataset labelled with at least two distinct representation schemes. Employing a single dataset, however, increases the risk of introducing biases and limitations that may undermine the validity and applicability of the study's conclusions due to some cultural or linguistic specifications. Therefore, in order to mitigate these concerns, we use two different datasets, both of which are labelled with the same two distinct representation schemes [11]. The datasets that meet the aforementioned criteria and we use in the study are:

- Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions [7]. Will refer to the dataset found in this study with "MADS" - Madrid Affective Database for Spanish.
- Turkish Emotional Word Norms for Arousal, Valence, and Discrete Emotion Categories [8]. Will refer to the dataset found in this study with "TEWN" - Turkish Emotional Word Norms.

Table 1: Datasets Representation Schemes and available ratings split by gender available

	Categorical	Dimensional	Per gender info
MADS	✓	✓	✓
TEWN	✓	✓	-

The affect representation schemes according to which both the datasets are labelled are *Categorical* and *Dimensional*.

Moreover, the datasets utilised in this study are constructed in different languages, which further enhances the generalisability of our findings.

In each of the studies, the used datasets consist of aggregated data entries, which represent the mean values derived from multiple individual ratings. The datasets do not provide access to personal or individual ratings. The aggregated form of the data ensures that individual ratings remain anonymous and confidential, preserving the privacy of all participants, however, having access to the mean values as well as standard deviation gives us enough information to perform our computations.

Additionally the MADS dataset has the supplementary breakdown of the average values according to gender which allows us to further explore generalisability.

2.2 Representation schemes

The two datasets have identical affect representation schemes, and these labelling schemes are as follows:

- **Categorical affect representation** refers to a way of categorizing or classifying emotions or affective states into *discrete categories or labels*. It involves representing emotions based on *predefined categories* or dimensions rather than representing them as continuous variables [12] [13].

  The specific categories contained in the the datasets are: Happiness, Anger, Sadness, Disgust and Fear.

- **Dimensional affect representation** refers to a method of representing emotions or affective states using *continuous dimensions*. Instead of categorizing emotions into discrete categories, dimensional affect representation focuses on capturing the underlying dimensions that describe the emotional experience [12] [13].

  The specific parameters contained in the datasets are Valence (negative - positive) and Arousal (calmness - action).

Furthermore, using datasets with identical labeling ensures consistency in the ground truth or reference labels across the datasets, enabling fair and meaningful comparisons between different models, algorithms, and approaches. By training and evaluating machine learning models on both datasets, any differences in performance can be attributed to the models themselves rather than discrepancies in labelling. It is worth noting that the experiments conducted to obtain these datasets were carried out in Spain and Turkey, using their respective languages, taking into account the influence of context and culture.

Considering the inherent characteristics of those two representation schemes, the translation from the Dimensional RS to the Categorical RS is the direction which aligns more naturally with the problem and is the one we will explore.

2.3 Pre-processing

Given the uniform data format observed in both MADS and TEWN datasets for each textual entry, several general observations can be made. Notably, all ratings associated with words in both datasets possess distinct numerical values for

Valence and Arousal, accompanied by their respective standard deviations. Therefore, the dimensional representation scheme values are explicitly provided, utilising the Mean Valence and Mean Arousal values.

Table 2: Representation schemes found in the datasets and used in the experiment with the added 'N' (Neutral) category

Categorical	Dimensional
$[H, A, S, F, D, N]$	(Valence, Arousal)
Singular label (exclusive)	$([0, 5], [0, 5])$

However, the categorical representation scheme presents information in the form of numerical values rather than categorical labels. Hence, a strategy similar to the approach described in the research papers [7] [8] is adopted. Specifically, any word with ratings falling within the mild range (not exceeding 2.5 for MADS and 50 for TEWN for all categorical entries) is assigned the neutral label 'N'. Conversely, all other words receive labels corresponding to the highest categorical rating they have obtained. To illustrate the finalised version of the representation schemes for each of the datasets, see Table 2. As well as to see the percentage of each categorical lable for each of the datasets, see Table 3.

Table 3: Distribution of labels per category of each of the two datasets, shown in percentages (%)

	Happiness	Anger	Sadness	Fear	Disgust	Neutral
MADS	35.1	9.6	9.9	13	3.4	29
TEWN	32.79	7.14	9.55	7.14	2.51	40.87

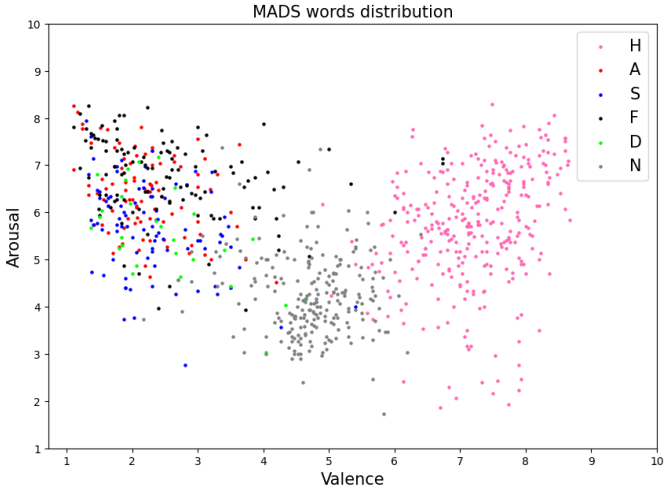

Figure 1: The distribution of the MADS dataset

2.4 Properties of chosen dataset

Further discussion regarding the features and specifications of the datasets used in the research is required.

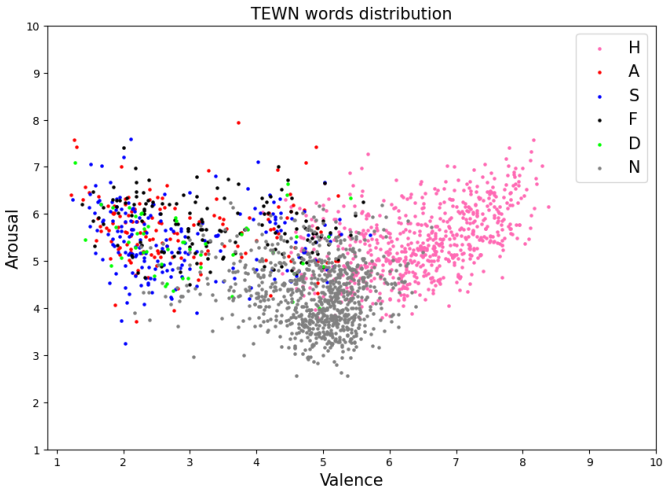

Figure 2: The distribution of the TEWN dataset

It is essential to address certain key points, such as the percentages of categorical representation for each category within the aforementioned datasets, see Table 3. Notably, there exists a clear imbalance in the distribution of percentages across the categories when exploring the categorical representation scheme. This may lead to a bias towards the majority classes and poor performance on minority classes (for example Disgust, which is the most under-represented in each of the datasets). Additionally, the scatter plot represents the distribution based on Valence and Arousal for MADS, Figure 3 and TEWN, Figure 4. It clearly points to the inbalance of the size of each category. Furthermore, upon comparing the distribution patterns depicted in the respective Figure 3 and Figure 4, it becomes apparent that the MADS dataset exhibits a higher degree of dispersion as opposed to the relatively concentrated distribution of the TEWN dataset, which exhibits overall fewer outliers in spite of its larger size.

Table 4: Size and information on raters in the study

	# words	# raters	# of women	# of men
MADS	875	660	507	153
TEWN	2031	1527	952	757

Furthermore, based on 4, the MADS dataset exhibits a notable gender bias as well, predominantly consisting of ratings provided by women. In contrast, the TEWN dataset demonstrates a relatively balanced distribution of ratings across genders. Such gender-based disparities in dataset composition may potentially impact the generalisability of cross-dataset translations, particularly if there exist significant differences in the perception of textual affect stimuli between men and women. Hence, in this paper, we will further investigate the extent to which gender influences the generalisability of our models [14].

2.5 ML Models

To address the task of translation between affect representation schemes, we can employ various supervised machine learning models. As our objective involves classifying entries into different categories, we have two choices: using a model that directly supports multi-class classification or adopting the One-vs-Rest (OvR) approach with a binary model. The following ML models were selected for this task, accompanied by a concise explanation for each choice:

Majority Classifier

A majority classifier serves as a simple baseline model that predicts the most frequent class within the training data for each input instance. It is commonly employed as a benchmark to assess the performance of more complex classification models. The majority classifier acknowledges the presence of significant imbalances in the dataset's class distributions, thereby reflecting the inherent bias in the data. By doing so, it helps to underscore the challenges posed by class imbalance for other models. Although the majority classifier represents a basic and sometimes simplistic approach, it plays a vital role in establishing a reference point for model evaluation and offers valuable insights into the effectiveness of advanced classification algorithms. During the evaluation process of classification models, each run and data split are also tested against the majority classifier as a reference point for the comparison with other models.

Logistic Regression

Logistic regression is originally designed for binary classification problems, however, we can extend the OvR strategy so that the model is applied to a multi-class problem.

Decision Tree

Decision Trees allow for multi-class classification, which is done in a transparent and intuitive way, as each node in the tree corresponds to a feature or attribute, and the branches represent the decisions based on that feature. This makes it easier to understand and interpret the mapping between labels. Therefore, for the specific task at hand it is a suitable machine learning model, as we can visually follow the choices it makes.

Random Forest

The Random Forest model is good for multi-class classification as it ensembles decision trees, which reduces over-fitting and improves the generalisation of the task. Furthermore, it handles high-dimensional data and automatically selects important features. Additionally, the Random Forest model provides interpretability through feature importance scores. Hence, it is a powerful and versatile classification model for multi-class problems.

Artificial Neural Network

The exact ANN we use is the MLPClassifier which is a relatively simple and straightforward example of an ANN. It allows us to optimise the activation function used in the modelling process, good at handling complex inputs.

3 Experimental Setup

The direction of the translation we explore is from Dimensional to Categorical representation scheme.

Table 5: Hyper-Parameters of the ML Models

Hyper-Param	Values	ML Model
Solver algorithm	['lbfgs','sag','saga']	Logistic regression
MaxDepth	[2, 3, 4]	Decision Tree
MaxDepth	[4, 5, 6]	Random Forest
Activation	['relu', 'logistic', 'tanh']	ANN

As mentioned before, in accordance to the studies from which the datasets were extracted, the representation schemes are as follows in Table 5.

We perform the training in 3 different steps:

- The first step is - *Singular dataset translation*.

  The training and testing procedures were conducted on a single dataset following the specifications provided in Table 5. All classifiers were executed concurrently, along with the Majority classifier, thereby ensuring consistency in data partitioning and reducing any potential biased influence on classifier performance. To mitigate bias and ensure comprehensive data representation, the classification process was be repeated 100 times and the mean, as well as, the standard deviations is presented. This method was applied to the MADS and the TEWN datasets separately. For the results see Figure 3 and Figure 4.

- The second step is - *Cross-dataset translation*.

  In this study, the model was trained on the MADS dataset and subsequently tested on the TEWN dataset, again complying with the specifications from Table 5. However, the approach was partially different as training and testing was performed on the entirety of each of the datasets. Similarly, the approach employed for training on the TEWN dataset and testing on the MADS dataset followed the aforementioned methodology.

- The second step is - *Cross-gender translation*

  Given that the gender information is solely available for the MADS dataset, our analysis primarily concentrated on this dataset for conducting cross-gender investigations. To facilitate this analysis, the labels within the Categorical representation scheme were recalculated to align with the respective categories of 'Female' and 'Male', thereby resulting in slightly differently labelled words from one another as well as the labelling which includes the mean labelling (*0.76* overlap between the male and female labels). We then proceed with training all the models on the male labels and testing on the female, while splitting the data in 80:20 and repeating 100 times in order to increase reliability. Similarly, the approach is mirrored for training on women ratings and testing on men.

4 Results

After following the methodology from the Experimental Setup section we obtain the following results:

Single dataset translation - MADS

We refer to Figure 3. The accuracy performance of all the ML Models is fairly consistent and all surpassing the accuracy of the majority classifier. The worst-performing model is Decision Tree overall. And the two best are ANN_tanh (*mean = 0.755, std = 0.019*) and Logistic Regression_sag and _saga (*mean = 0.751, std = 0.018*).

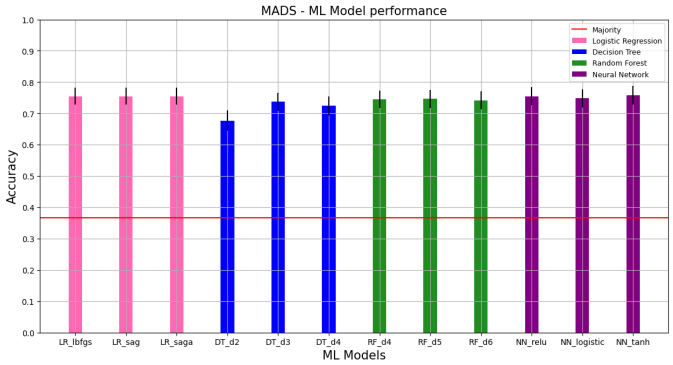

Figure 3: MADS: Distribution of ML Models' performance on a single dataset, 100 Epochs, 80:20 split

Moreover, our analysis extends to the evaluation of the overall proficiency in accurately predicting each label within the Categorical representation scheme. We employ the ANN_tanh ML model for this evaluation, as depicted in Table 6. Evidently, the categories that are encountered most frequently exhibit the highest levels of predictive accuracy (notably, H and N with f1-scores *0.961* and *0.885* respectively).

Table 6: MADS: Mean of 100 epochs classification reports using an Artificial Neural Network ('tanh'), 80:20 split.

	H	A	S	F	D	N
precision	0.958	0.416	0.358	0.555	0.0	0.864
recall	0.965	0.236	0.548	0.652	0.0	0.913
f1-score	0.961	0.289	0.43	0.593	0.0	0.885
support, in percentage	36.74	10.86	8.69	13.37	3.83	26.51

Singular dataset translation - TEWN

We refer to Figure 4. Again, similar to the MADS dataset, the accuracy performance of all the ML Models is fairly consistent over the TEWN dataset. All of the models surpass the accuracy of the majority classifier. The worst-performing model is the Decision Tree overall. However, here the best one is Logistic Regression with all of its hyper-parameters (*mean = 0.713, std = 0.021*).

Moreover, our analysis extends to the evaluation of the overall proficiency in accurately predicting each label within the Categorical representation scheme. We employ the LR_saga

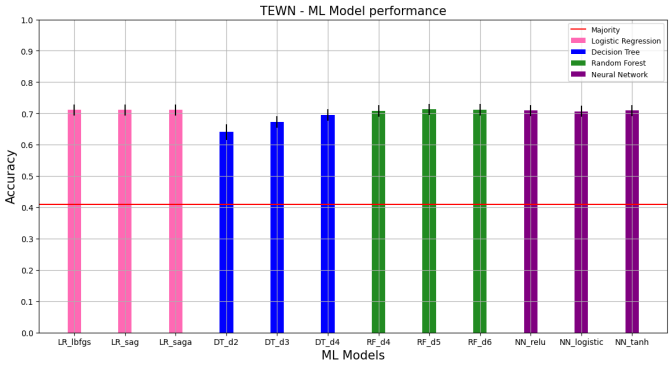

Figure 4: TEWN: Distribution of ML Models' performance on a single dataset, 100 Epochs, 80:20 split

ML model for this evaluation , as depicted in Table 7. Evidently, the categories that are encountered most frequently exhibit the highest levels of predictive accuracy (notably, N and H, wiht f1-scores of *0.822* and *0.859* respectively).

Table 7: TEWN: Mean of 100 epochs classification reports using Logistic Regression ('saga'), 80:20 split.

	H	A	S	F	D	N
precision	0.869	0.289	0.379	0.315	0.0	0.772
recall	0.851	0.092	0.578	0.201	0.0	0.879
f1-score	0.859	0.121	0.452	0.237	0.0	0.822
support, in percentage	32.76	7.17	9.60	7.13	2.54	40.78

Cross-dataset translation

Here we present the performance of ML models when trained on one dataset and subsequently tested on the other. The graphical representation of this analysis can be seen in Figure 5. It is worth noting all of the models outperform their respective Majority classifier counterpart. However, it is noteworthy that the optimal ML model varies depending on the direction of the analysis. For the case of the training on MADS and testing on TEWN, the best ML model is the ANN (*0.714*). Whereas the case of the model trained on MADS and tested on TEWN the best ML model is LR (*0.735*).

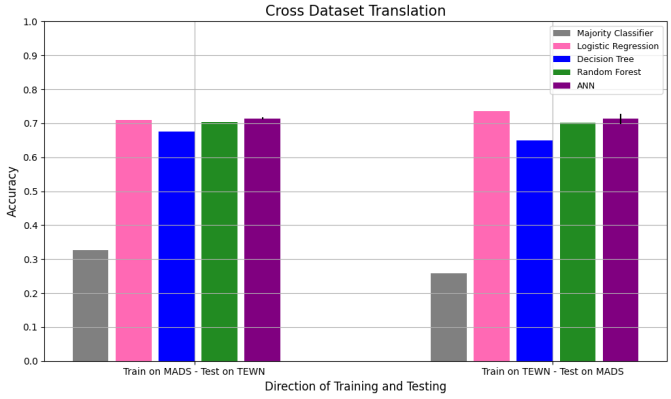

Figure 5: MADS & TEWN: Cross-dataset translation

Additionally, there exists a substantial disparity between the accuracy achieved between the Majority classifier and that achieved by the best performing ML model between the two runs. This discrepancy signifies a significant enhancement provided by the ML model, showcasing its notable superiority over the baseline Majority classifier in both of the directions, but especially when the model is trained on TEWN and tested on MADS.

Cross-gender translation

Given the limited availability of gender-specific ratings, this experiment is restricted to using the MADS dataset.

When looking at Figure 6 and Figure 7 we again see that all the models across each of the directions surpass the Majority classifier accuracy in their respective scenario. The performance of ML models remains fairly consistent across both scenarios, however, in comparison to the performance of the ML models in the single dataset translation in Figure 3, the accuracy is around 0.1 lower for each of the models, other than the Majority classifier.

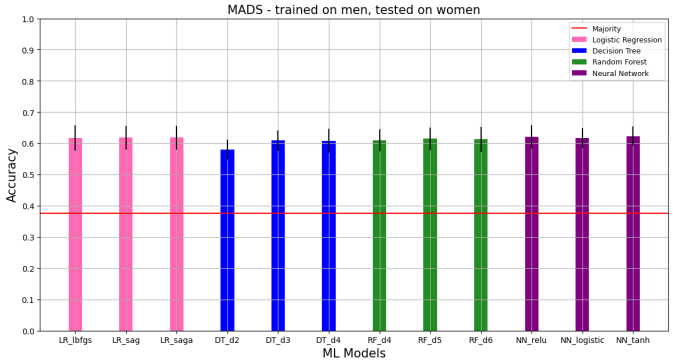

Figure 6: MADS - trained on men, tested on women

Regarding the comparison of accuracies between the two translation scenarios, it can be concluded that, overall, the ML models trained on the female labels and tested on the male labels exhibit slightly better performance than their counterparts trained on male and tested on women, for each respective ML model.

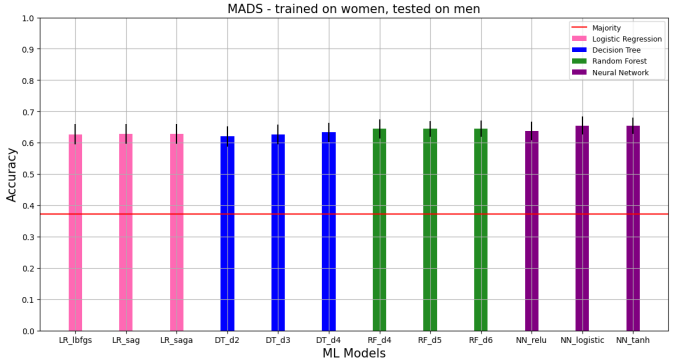

Figure 7: MADS - trained on women, tested on men

5 Discussion

In this study, we conducted an analysis of Machine Learning Mmodels for translation between Affect Representation Schemes (Dimensional to Categorical) using the MADS and TEWN datasets.

In all of the aforementioned scenarios, our analysis showed that the f1-score was consistently highest for two sentiment labels, namely 'H' and 'N'. A potential explanation for this observation is that these two labels are the most frequently encountered in each of the datasets. However, it is important to consider the additional factor that may contribute to their higher f1-scores, namely their distinct mean Valence and mean Arousal values when compared to the other four labels.

In support of this claim, we present the mean Valence and Arousal values for each sentiment label in Table 8 and Table 9, respectively. It is evident that the H' and N' labels exhibit significantly different mean Valence and Arousal values compared to the other sentiment labels. Concretely, label 'H' tends to have a higher mean Valence value, indicating a more positive sentiment, while label 'N' has a relatively neutral Valence value. In terms of Arousal, the 'N' label shows the lowest mean value, which indicates a mild reaction to the respective word.

Table 8: MADS: Valence & Arousal, mean & standard deviation

	H	A	S	F	D	N
Valence Mn	7.21	2.22	2.32	2.62	2.51	4.73
Valence SD	0.75	0.64	0.73	1.12	0.86	0.68
Arousal Mn	5.71	6.37	5.65	6.73	5.51	4.24
Arousal SD	1.24	0.84	0.96	0.85	0.96	0.86

Table 9: TEWN: Valence & Arousal, mean & standard deviation

	H	A	S	F	D	N
Valence Mn	6.48	2.90	2.93	3.38	3.04	4.86
Valence SD	0.88	1.07	1.09	1.08	1.07	0.70
Arousal Mn	5.43	5.65	5.42	5.75	5.37	4.29
Arousal SD	0.71	0.68	0.70	0.57	0.65	0.70

Furthermore, the distinct mean Valence and Arousal values observed for the 'N' make it possible for the model to accurately predict them. On the other hand, the relatively similar mean values of the labels 'A', 'S', 'F', 'D' pose as a challenge for the ML Models in providing an accurate prediction.

To address this challenge and likely improve the performance of the translation the introduction of a third dimension to the Dimensional Representation Scheme [15] could be explored. By incorporating an additional dimension, such as Dominance, the model could capture more nuanced variations in sentiment expression, as currently the model does not distinguish well the "negative" emotions.

The observed differences in the extremity of the mean Valence and mean Arousal values between datasets in cross-

dataset translation may be attributed to cultural disparities, as well as the difference in input words. Cultural influences can impact the expression and perception of emotions and language and therefore this can greatly impact the affect ratings [16]. However, another point is the fact that the datasets contain different words, which may skew the average ratings of the datasets [17].

Table 10: MADS: Some chosen words which were labelled differently by women and men

Word	Labelled by women	Labelled by men
to admire	D	H
to harass	F	A
to threaten	F	A
envious	D	A
to infect	D	F
embarrassment	A	F

In the cross-gender translation scenario, it is important to note that there is an initial overlap of only 0.76 between the extracted labels of the two genders. Table 10 presents a selection of words that evoke different categorical emotions. The reason for still achieving an accuracy over 0.6 in this scenario may be attributed to the alignment of these differences with the Dimensional Representation scheme. This alignment enables the preservation of the relationship between emotions, hence facilitating accurate translation despite the variation in labels across the two genders.

6 Conclusion

In this study, we analysed different Machine Learning Models for translating between Affect Representation Schemes using the MADS and TEWN datasets.

The Categorical emotion labels 'H' and 'N' consistently achieved the highest accuracy scores. Their frequent occurrences in the datasets and distinct mean Valence and Arousal values likely contribute to this performance. The mean Valence and Arousal values support these findings, showing that the 'H' label had a more positive Valence and that label 'N' exhibited a relatively neutral Valence with low Arousal. Accurate predictions were facilitated by the distinct values of the labels 'H' and 'N', while the similarity of 'A', 'S', 'F', and 'D' labels posed a challenge for all tested ML Models.

To improve translation, we can try incorporating a third dimension, such as Dominance into the Dimensional Representation Scheme to capture more nuanced variations in the emotional expression, especially for the "negative" emotions.

Differences in mean Valence and Arousal values between datasets may stem from cultural disparities and variations in input words, impacting affect ratings. However, the ML Models generalised fairly well, without a significant reduction in accuracy.

In cross-gender translation, despite initial labelliing differences, accuracies above 0.6 were achieved due to the alignment of variations with the Dimensional Representation scheme, preserving the relationship between emotions.

Overall, the translation between different Affect Representation Schemes is feasibly for distinguishing between 'Negative', 'Neutral' and 'Positive' words ('A,S,D,F', 'N', 'H'), as well as generalise between different datasets and genders. However, further work is required to establish a robust and reliable translation framework.

7 Responsible Research

The conducted experiment and methods explored in this research project conform to the responsible values and pillars of high academic standards laid out by the TU Delft. The displayed results are completely transparent and objective, in line with the development process. No changes have made to tamper with the results or change the meaning of the findings.

As we are using datasets of previously carried out research, we have not obtained any new raw data and therefore not had any moral or ethical concerns in the collection process. All the data used is cited accordingly any we do not claim any rights to the raw information.

8 Future work

This research has certain limitations that should be acknowledged. Firstly, the absence of individual ratings restricts the exploration of specific variations and nuances within the dataset. However, despite this limitation, the aggregated data provides a valuable representation of overall rating tendencies, enabling broader conclusions and informed decision-making.

Additionally, the uneven distribution of items across labels introduces a potential bias in the training and evaluation of the translation model. Labels with a higher number of items, such as Happiness and Neutral, received more emphasis during training, potentially leading to better performance on those specific labels. Conversely, labels with fewer items may be underrepresented in the training data, posing challenges for accurate translation within those categories. It is crucial to consider this disproportionate distribution when interpreting the accuracy of automatic translation, as the performance of the model may be influenced by label distribution, and the accuracy scores may not equally reflect translation quality across all categories.

To address the issue of disproportionate labeling, various strategies can be explored, including data augmentation techniques, re-sampling methods, or incorporating weightings during training. These approaches aim to balance the representation of items within each category, ensuring a fair and unbiased evaluation of translation performance across all labels.

To ensure the cross-dataset generalisation and address potential biases, collecting ratings from diverse cultures and languages for the same dataset would be valuable. This would help evaluate the impact of cultural and linguistic variations on affect ratings and ensure the reliability of the machine learning approach. Collecting additional datasets that include gender-specific ratings would be instrumental in addressing biases and ensuring a more comprehensive analysis. These datasets would contribute to the identification and mitigation

of potential gender biases, thereby enhancing the overall accuracy and fairness of the translation process.

9 Acknowledgements

I would like to extend my appreciation to my research supervisors - Postdoctoral Researcher Bernd Dudzik and Assistant Professor Chirag Raman for their guidance and constructive suggestions during the planning and development of this research. Their expertise has been crucial in shaping the trajectory of this Research Project. I am equally thankful to my colleagues for their valuable contributions and collaborative efforts - Alissia Rugina, Ivan Dimitrov and Shuang Liu.

References

[1] Nourah Alswaidan and Mohamed El Bachir Menai. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987, August 2020.

[2] Klaus R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, December 2005.

[3] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. Survey on sentiment analysis: evolution of research methods and topics. *Artificial Intelligence Review*, January 2023.

[4] Hans Stadthagen-González, Pilar Ferré, Miguel A. Pérez-Sánchez, Constance Imbault, and José Antonio Hinojosa. Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness. *Behavior Research Methods*, 50(5):1943–1952, October 2018.

[5] Dale J. Cohen, Katherine A. Barker, and Madeline R. White. A standardized list of affect-related life events. *Behavior Research Methods*, 50(5):1806–1815, October 2018.

[6] Agnieszka Landowska. Towards New Mappings between Emotion Representation Models. *Applied Sciences*, 8(2):274, February 2018.

[7] J. A. Hinojosa, N. Martínez-García, C. Villalba-García, U. Fernández-Folgueiras, A. Sánchez-Carmona, M. A. Pozo, and P. R. Montoro. Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, 48(1):272–284, March 2016.

[8] Aycan Kapucu, Aslı Kılıç, Yıldız Özkılıç, and Bengisu Sarıbaz. Turkish Emotional Word Norms for Arousal, Valence, and Discrete Emotion Categories. *Psychological Reports*, 124(1):188–209, February 2021.

[9] Lisa M. Bauer and Jeanette Altarriba. An Investigation of Sex Differences in Word Ratings Across Concrete, Abstract, and Emotion Words. *The Psychological Record*, 58(3):465–474, July 2008.

[10] Geneva M. Smith and Jacques Carette. What Lies Beneath—A Survey of Affective Theory Use in Computational Models of Emotion. *IEEE Transactions on Affective Computing*, 13(4):1793–1812, October 2022.

[11] Joost Broekens and Willem-Paul Brinkman. AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies*, 71(6):641–667, June 2013.

[12] Rafael A. Calvo and Sunghwan Mac Kim. EMOTIONS IN TEXT: DIMENSIONAL AND CATEGORICAL MODELS. *Computational Intelligence*, 29(3):527–543, August 2013.

[13] M. Horvat, A. Stojanovic, and Z. Kovacevic. An overview of common emotion models in computer systems. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1008–1013, Opatija, Croatia, May 2022. IEEE.

[14] Yaling Deng, Lei Chang, Meng Yang, Meng Huo, and Renlai Zhou. Gender Differences in Emotional Response: Inconsistency between Experience and Expressivity. *PLOS ONE*, 11(6):e0158666, June 2016.

[15] Iris Bakker, Theo Van Der Voordt, Peter Vink, and Jan De Boon. Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. *Current Psychology*, 33(3):405–421, September 2014.

[16] W. Jiang. The relationship between culture and language. *ELT Journal*, 54(4):328–334, October 2000.

[17] Margaret M. Bradley and Peter J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, March 1994.