Utrecht University

Department of Information and Computing Science

Applied Data Science master thesis

Screening narratives of adolescents: A comparative analysis of BERTje, tf-idf, and multilingual DistilUSE feature extractors

First examiner:
Rens van de Schoot
Laura Hofstee

Second examiner:
Dr. Ayoub Bagheri

Candidate:
Kajol Atwani
(8059446)

In cooperation with:
AsReview

July 2, 2023

Abstract

In psychological research, understanding identity formation involves screening and labeling textual data, such as narratives by teenagers about turning points in their lives. This process is similar to systematic screening for meta-analyses. Manual screening can be enhanced with AI-aided tools. This study explores active learning for analyzing Dutch narratives of teenagers' turning points. A reliable feature extractor is needed to capture psychological, structural, and content dimensions in Dutch narrative identity data. The effectiveness of BERTje, tf-idf, and DistilUSE is compared for analyzing narrative identities among Dutch adolescents. BERTje performs best, demonstrating superior recall values and reduced average time to discover (ATD). BERTje shows promise for analyzing narrative data and extracting insights on identity formation. The performance gain may differ from screening abstracts. Showcasing BERTje's effectiveness and the potential of AI-aided screening tools offer insights for comparative analysis and research on identity narratives. Integrating AI-aided tools in developmental psychology supports informed decision-making and enhances understanding of narrative identities in interventions. These findings contribute to utilizing AI-aided tools in data analysis, understanding human experiences, and advancing psychological research and interventions.

Keywords: *Narrative Analysis*, *Feature Extractors*, *BERTje*, *DistilUSE*, *tf-idf*, *Active Learning*, *AI-aided Screening*

Contents

1. Introduction

Adolescence is a transitional period marked by substantial changes in physical maturation, cognitive abilities, and social interactions (Laird, 2013). Identity formation is thought to be the key developmental task of adolescence (Erik H Erikson, 1950). During this period, individuals gain cognitive capacities for abstract thinking and begin searching for sameness and continuity of the self (Klimstra, 2013). The development of identity, as proposed by E. H. Erikson (1968), is a continuous journey that becomes particularly important during adolescence. Young Dutch adolescents, around the age of 14-15, face societal expectations, including making decisions regarding their secondary school curriculum, which have a major impact on their identity development (Doeselaar et al., 2020).

Researchers in developmental psychology have extensively explored narrative identities, with a specific emphasis on the pivotal moments that shape adolescents' lives. 'Project Me' (Tilburg University, 2023), a notable research endeavor in this area, engaged a significant number of actively participating adolescents. These young participants shared the turning points in their lives, and the collected texts were subsequently coded for analysis. The project provides valuable insights into the developmental aspects of narrative identities and their impact on the lives of adolescents.

However, the current manual process of screening and labeling narratives in the context of identity formation and psychological interventions is time-consuming and inefficient. There is a potential for improvement by leveraging AI-aided tools, such as ASReview (ASReview team, 2023), to accelerate the systematic review process. Adopting an AI-aided tool for systematic review allows for efficient identification of themes related to psychological, structural, and content dimensions in the data, comparable to manual human tagging (Van den Brand and Schoot, 2021). This approach not only saves time but also ensures consistent and reliable results. Therefore, this research aims to explore the application of AI-aided tools in the analysis of narrative identities, offering a promising solution to existing challenges and highlighting the potential of AI-aided tools in enhancing the understanding of identity narratives and supporting psychological interventions.

Narrative approaches operationalize adolescents' identity formation by focusing on the construction of autobiographical life stories and the formation of identity commitments (Doeselaar et al., 2020). According to Singer (2004), narrative identity represents a selective and subjective account of how one came to be the person one currently is. Autobiographical memories that are significant to individuals can be woven into an extended story that defines the self (D. P. McAdams, 1993). The personal narrative plays a crucial role in developmental psychology, allowing individuals to experience a profound sense of integration, meaning, and purpose. According to Dan P McAdams and Kate C McLean, narrative coding encompasses four general categories: motivational themes, affective themes, structural elements, and themes of integrative meaning.

Motivational themes focus on an individual's goals and strivings. In the context of 'Project Me', agency is one of the motivational themes actively sought out in the narratives. Agency refers to the degree to which the protagonist can initiate changes, exert control over their experiences, and impact their own life (Adler et al., 2017). Motivational themes, along with autobiographical reasoning, have been identified as crucial indicators in these narratives (Kate C McLean et al., 2020). Affective themes, on the other hand, refer to the emotionality of a narrative, such as positive or negative valence (Kate C. McLean and Syed, 2020). They provide insights into the current and future well-being of an individual, including the theme of change in emotional state over the course of the narrative, such as redemption stories.

Structural elements, the third category, refer to the overall coherence of the narrative presented by the adolescent, which is generally associated with the psychological health of the individual. Autobiographical reasoning involves creating a coherent and integrated life story by linking personal experiences and aspects of the self (Habermas and Reese, 2015). Self-event connections are theorized to be highly important for the development of a coherent narrative identity and considered adaptive in nature (Pasupathi, Mansour, and Brubaker, 2007). Individuals who struggle to articulate their experience in terms of characters, ordered events, context, and specific details tend to exhibit poorer overall functioning (Kate C. McLean and Syed, 2020). Lastly, themes of integrative meaning refer to the connection of events the narrator makes to the self rather than the recollection of event details. The meaning-making process plays a crucial role in examining the development of identity, where individuals utilize past experiences to gain insight into and define their sense of self. This theme is associated with the psychological well-being of an individual (Kate C. McLean and Syed, 2020).

The coding of narratives involved a three-step process to ensure accurate analysis. First, the coding manuals were adapted to suit the current data. Next, research assistants were trained to apply the coding systems to the new narratives. Lastly, the narratives were coded by trained coders to ensure consistency and avoid code drift.

Before applying machine learning techniques, AI-aided tools require the processing of text data into numerical representations known as vectors. This step is crucial for effectively analyzing and extracting insights from textural data, including the narratives written in Dutch by adolescents for 'Project Me'. The three feature extractors that will be compared are BERTje, tf-idf, and multilingual DistilUSE. BERTje, being a monolingual feature extractor, focuses solely on Dutch and captures its unique nuances. Tf-idf (default) calculates the importance of terms in a document relative to the collection of documents, capturing their relative significance. The multilingual feature extractor leverages language-specific knowledge and resources to effectively analyze Dutch narratives. Specifically, based on the knowledge that BERTje was fine-tuned exclusively for Dutch NLP tasks (De Vries et al., 2019), it is expected that BERTje will outperform other feature extractors in capturing the nuances and features of Dutch narratives written by adolescents. In this paper, the use of BERTje, tf-idf, and multilingual DistilUSE feature extractors will be compared to identify the best feature extractor for screening narratives in the Dutch language written by adolescents using active learning.

Overall, this research strives to enhance our understanding of narrative identities and their role in adolescence by exploring the potential of AI-aided tools in the systematic review of narratives. By establishing an effective feature extractor for Dutch narratives, we can pave the way for psychologists and researchers to leverage AI-aided tools to streamline the screening phase and gain valuable insights into the developmental processes of identity formation. Additionally, this study highlights the need for further research to optimize the pipeline and maximize the benefits of active learning tools in the field of narrative analysis.

2. Methodology

The methodology section presents a comprehensive overview of the research approach and techniques employed in this study. It begins by providing a detailed description of the data, including its source and any necessary preparations. The section then discusses the simulation setup, highlighting the specific metrics utilized, which will be further elaborated upon in the results section. Additionally, it delves into the feature extraction process, offering in-depth insights into the implementation of three key techniques: TF-IDF, BERTje, and Multilingual DistilUSE. While TF-IDF is described, the methodology further explores the pre-training and fine-tuning processes of BERTje and Multilingual DistilUSE feature extractors, which were utilized in this study.

2.1 Description of the data

The dataset used to evaluate the performance of different feature extractors is from 'Project Me,' coordinated by Tilburg University (Tilburg University, 2023). The data for 'Project Me' was collected through an hour-long questionnaire discussing turning point narratives written in Dutch as described by 1580 Dutch second and third-year students of different secondary schools during 2015 and 2016. Psychologists systematically analyze these written accounts of adolescents expressing their emotions to identify any instances of atypical behavior. The participants of the study were from all three educational levels in the Netherlands: pre-vocational education (18.2%)(MAVO), higher secondary education (38%)(HAVO), and preparatory scientific education (43.8%)(VWO). (Doeselaar et al., 2020)

According to Doeselaar et al. (2020), out of the 1941 adolescents who participated, 50 incomplete narratives were excluded from the study. Out of these students, 311 could not come up with a turning point in their lives, and others who did not want to disclose their turning points were also excluded from the study, leaving 1580 participants. The majority of the participants are female (56.2%), and their average age is 14.7 years old (Doeselaar et al., 2020). The explicit instructions for the turning points included what happened, when it happened, who was involved, what they were thinking and feeling, why the experience was significant, and what it could say about them and their personality (Doeselaar et al., 2020).

The data has three types of labels: psychological, structural, and content dimensions as mentioned earlier. Labels such as 'event_negvalence', 'event_posvalence', and 'agency' are psychological dimensions that refer to the interpretation of the event by the adolescents as well as their emotions behind a story. Agency, as defined in the introduction, refers to the extent to which the protagonist can initiate changes, exert control over their experiences, and impact their own life (Adler et al., 2017). These psychological dimensions represent the motivational (agency column) and affective themes categories in narrative coding. The structural theme category of narrative coding is synonymous with the structural dimension. The structural dimension labels are applied when the text is a story that is explicitly linked to the adolescents themselves, for example, the 'w1N_selfEventconnections'. The labels with 'event_content' in their name are content dimensions which are when the story is about a turning point in the adolescent's own life. The content dimensions are synonymous with themes of the integrative category in narrative coding. It is important to note that a narrative can exhibit multiple dimensions within these categories.

Psychology students trained with a narrative coding guide assign binary values to these columns. The value '1' was assigned to a column if the narrative depicted the theme the column represented, while the value '0' was assigned if the narrative lacked the theme. They apply their knowledge to a dataset with example narratives and then discuss the labels they tagged with the group. The Principal Investigator (Theo Klimstra) made the final decision for the dataset labels and explains their correctness or incorrectness. The benchmark for agreeability for a label was around 70% agreements (Van Doeselaar, 2019). Calibration meetings were held after labeling a batch of around 200 texts.

2.2 Preparation of the data

The original unmodified dataset from 'Project Me' had 27 columns or characteristics. During the preparation of the data, it was discovered that three of these columns, which had characteristic names starting with 'Coh', contained noisy labels due to improper data entry. This issue required special attention to ensure the accuracy and reliability of the dataset. A '0' value was entered when the condition was not satisfied or when the data was unknown (NA). These were the 'Coh_Context_FinalBinary', 'Coh_Chronology_FinalBinary', and 'Coh_Theme_FinalBinary' columns that were removed.

The 'Self_content_other_Binary' column was also omitted because there were no '1' values for any of the adolescent participants, as evident from figure 2.1a.

Among the 1588 participants, 81 individuals had missing values for certain columns, including the field containing their narratives describing turning points. As mentioned in 'Adolescents' Identity Formation: Linking the Narrative and the Dual-Cycle Approach,' these participants should be omitted from the dataset received from 'Project Me' (Doeselaar et al., 2020). After discussing with one of the researchers, Jaap Denissen from 'Project Me,' the participants with missing observations were removed, resulting in 1507 participants' having complete data. The last step of the data preparation consisted of creating separate datasets for each column of the 'Project Me' dataset.

2.3 Simulation Set-up

ASReview, an open-source AI-aided learning tool, was employed to streamline the screening and systematic labeling of textual data, significantly reducing screening time by up to 95%(ASReview developers, 2023). ASReview's MakItA (MAKe IT Automatic) is a workflow generator designed for simulation studies, automating the creation of study frameworks and

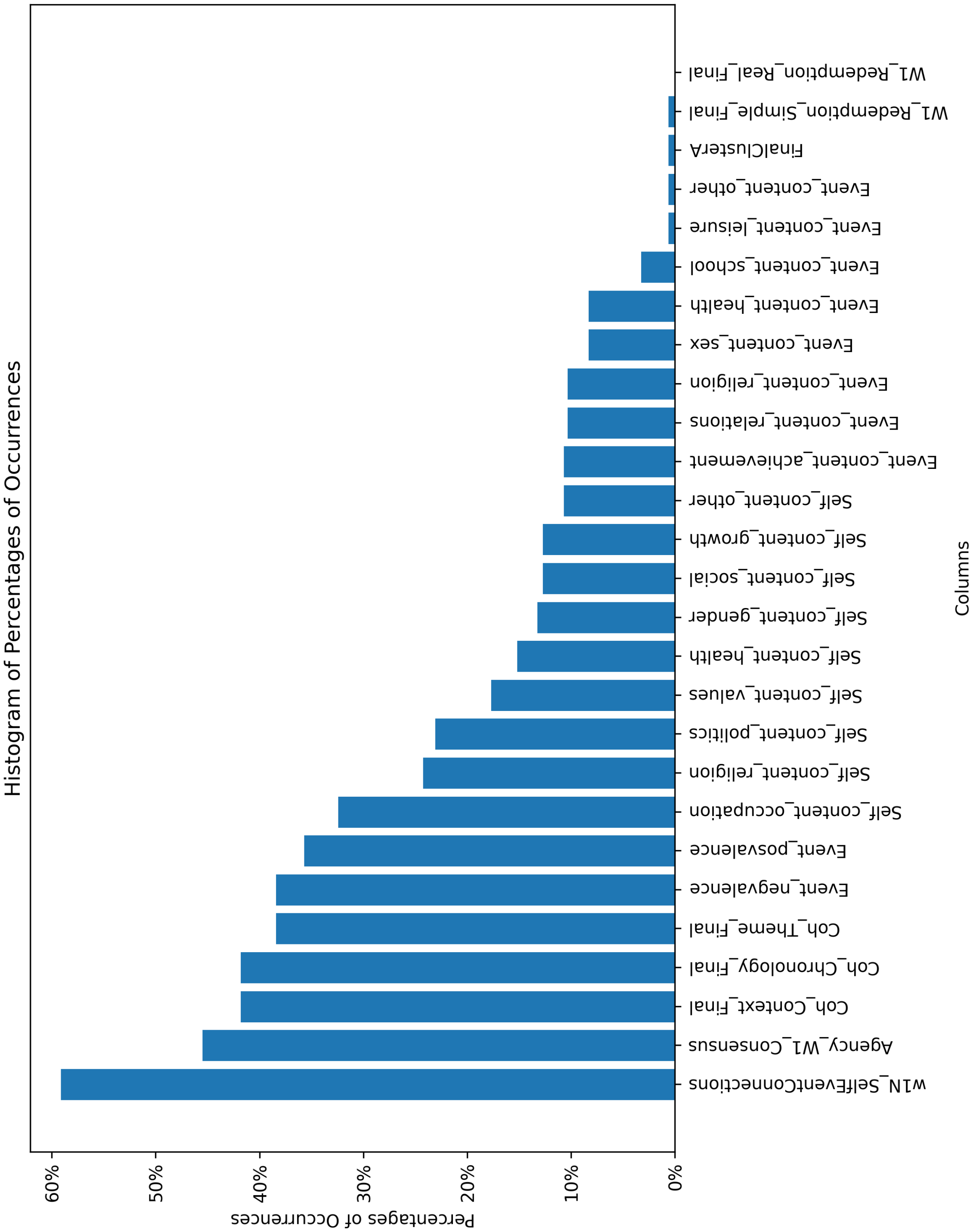


Figure 2.1: Histogram of Percentage of Ones

(a) Shows the spread of the percentage of occurrences, representing the presence of the themes represented by the columns in the narratives.

simplifying the process (Jelle Teijema et al., 2023). Simulations mimic the human screening process in combination with an active learning model, enabling the evaluation of model performance using various metrics. To analyze and visualize simulation results, ASReview provides the insights package (developers, n.d.), offering statistical results and plotting capabilities for metrics like recall, Work Saved over Sampling (WSS), and average time to discover (ATD).

The BERTje (Vries et al., 2019) and DistilUSE (Reimers and Gurevych, 2019a) models were obtained from HuggingFace and integrated into ASReview version 1.2 via a template(J. Teijema, 2021) (J. Teijema, 2023). The code for reproducibility is available on GitHub (Atwani, 2023), while the data used in the simulation is not publicly accessible information to obtain data on GitHub. Each column was processed separately using the multi-model MakItA (version 0.6.3) template, combining different feature extractors with logistic regression, naive Bayes, random forest, and support vector machine models.

2.4 Metrics Used

To assess the performance of the different feature extractors and models, various metrics were used. Recall plots were generated for each column, presenting recall values for each feature extractor-model combination. Additionally, Excel files were generated for each column, containing metrics such as recall at different percentages, work saved over sampling (WSS), the average time to discover (ATD), and ERF (Extra Relevant records Found).

The metrics used to compare the different feature extractors are from the insights package available on ASReview (developers, n.d.). Recall, also known as relevant records found (RRF), represents the proportion of relevant records identified in a simulation at specific set percentages (developers, n.d.). ATD, or average time to discover, represents the average number of records that need to be screened to identify all relevant records within the dataset (Ferdinands et al., 2020).

The analytic strategy in this study focuses on evaluating recall values and ATD for each column in the dataset. Recall values provide insights into the effectiveness of models and feature extractors in retrieving relevant records within each column. By comparing recall values, it is possible to assess the performance of different models and determine the best-performing one. The ATD metric helps evaluate the efficiency of models and identifies the feature extractor that minimizes screening effort. The WSS was not considered in the analytic strategy as recall and ATD captures the insights needed for this study.

2.5 Feature Extractors

The three feature extractors used in this paper, the tf-idf, monolingual BERTje, and distilUSE multilingual feature extractors are extensively described in the subsections below. Additionally, to compare the performance of these extractors across different datasets and dimensions, a box plot analysis was conducted. The box plots represent various metrics, including recall at different thresholds (0.1, 0.25, 0.5, 0.75, and 0.9), WSS at the 0.95 quartiles, ERF at the 0.1 quartiles, and ATD for each feature extractor. These box plots provide insights into the distribution and consistency of performance for each feature extractor, as discussed further in the Results section.

2.5.1 Tf-idf

Tf-idf (term frequency-inverse document frequency) is the default feature extractor used in ASReview for conducting simulations. It plays a crucial role in the AI-aided learning tool by evaluating the relevance of words in a document collection(Stecanella, 2019). Tf-idf assigns a score to each word based on its frequency in the document (TF) and its rarity in the corpus (IDF), indicating its importance to the specific document (2.1). These TF-IDF scores serve as representations of each document and are instrumental in assessing document relatedness and facilitating the AI-powered learning process.

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D) \tag{2.1}$$

$$\text{TF}(t, d) = \frac{\text{number of occurrences of term } t \text{ in document } d}{\text{total number of terms in document } d} \tag{2.2}$$

$$\text{IDF}(t, D) = \log\left(\frac{\text{total number of documents in the corpus } D}{\text{number of documents containing term } t}\right) \tag{2.3}$$

There are two statistical methods used in tf-idf: term frequency and inverse document frequency. Term frequency (TF) is the total number of occurrences of a term in a document against the total number of all words in the document (2.2). The inverse document frequency (IDF) measures the weight of a given word in the entire document. Inverse document frequency refers to how rare or common a word is, which depicts the importance of the word in the document to provide information (2.3)(Q. Liu et al., 2018).

To apply tf-idf, the dataset is first tokenized, breaking it down into individual words or terms. Then, the term frequency (TF) of each term is calculated, followed by the computation of the inverse document frequency (IDF). The TF and IDF values are combined to calculate the TF-IDF score for each term, providing a measure of its importance. These TF-IDF scores can be used as representations of each document. To assess the relatedness of documents, cosine similarity is applied to measure the similarity between the TF-IDF vectors (Analytics Vidhya, 2021).

2.5.2 monolingual BERTje

BERTje is a monolingual Dutch BERT model based on the transformer-based pre-trained language model BERT. It shares the same architecture and parameters as BERT and has been specifically trained on a large and diverse dataset of 2.4 billion tokens (De Vries et al., 2019). BERT, which stands for Bidirectional Encoder Representations from Transformers, has achieved state-of-the-art results in various natural language processing (NLP) tasks such as question answering, sentence classification, and sentence-pair regression (Devlin et al., 2018; Vaswani et al., 2017).

The performance of BERTje has been compared with multilingual BERT on word-level NLP tasks. According to De Vries et al. (2019), BERTje consistently outperforms multilingual BERT, although the performance comparison may vary depending on the specific task and evaluation metric.

During pre-training, BERTje was trained on high-quality cleaned Dutch text, from multiple corpora. The pre-training dataset included contemporary and historical fiction novels, the SoNaR-500 multi-genre reference corpus (Oostdijk et al., 2013), TwNC (a multifaceted Dutch News corpus) (Ordelman et al., 2007), and web news articles from four Dutch news websites. The Wikipedia dump from October 2019 was also included. Overlapping material was removed from the pre-training dataset (De Vries et al., 2019).

BERTje was pre-trained on two objectives: Next Sentence Prediction (NSP) and Masked Language Modeling (MLM) (Devlin et al., 2018). NSP encourages the model to learn the semantic coherence between sentences, while MLM focuses on embedding words based on their context. For BERTje, the MLM task was modified to mask consecutive words instead of random words, to ensure the accurate embedding of unmasked words. Figure 2.2 shows that pre-training procedures for BERT and BERTje are similar, but the fine-tuning procedures for BERTje are specifically tailored for Dutch NLP tasks (Devlin et al., 2018).

After pre-training, BERTje undergoes fine-tuning to adapt it to perform well on specific tasks. In the paper by De Vries et al. (2019), the pre-trained

BERTje model is fine-tuned on annotated data from Dutch CoNLL-2022, Lassy Small treebank, and the SoNaR-1 corpus. These datasets provide information on named-entity recognition, part-of-speech tagging, sentiment analysis, and other classification tasks relevant to the Dutch language respectively (De Vries et al., 2019; Sang and De Meulder, 2003; Van Noord et al., 2013; Delaere, Hoste, and Monachesi, 2009).

2.5.3 DistilUSE multilingual feature extractor

The distilUSE multilingual model is a language model specifically designed for sentence embeddings, which maps sentences to a dense vector space of 512 dimensions. It utilizes a transformer encoding model that incorporates bidirectional self-attention to compute context-aware representations of tokens within a sentence. By considering both the ordering and identity of tokens, the model generates a meaningful sentence-level embedding (Vaswani et al., 2017).

To train the distilUSE multilingual feature extractor, various datasets from the OPUS website were used (Tiedemann, 2012). These datasets encompassed political sources like Europarl, NewsCommentary, and UNPC, as well as media sources like JW300, OpenSubtitles2018, and TED2020 (Koehn, 2005; Ziemski, Junczys-Dowmunt, and Pouliquen, 2016; Agic and Vulic, 2019). Additionally, datasets such as Wiki-Matrix (consisting of sentences from Wikipedia, often of lower quality) and Tatobe (a large database of example sentences and their translations) were utilized for training (Schwenk et al., 2019; Reimers and Gurevych, 2020).

The distilUSE model is based on the idea of distilmBERT, a student model emulating the behavior of the monolingual mUSE model, to minimize mean-square loss. Figure 2.3 illustrates how the distilUSE model aligns sentence embeddings, while being trained on parallel translated sentences (Yang et al., 2019; K. Liu, Wang, and Zhang, 2022; Reimers and Gurevych, 2020).

Two key properties of the distilUSE multilingual model are its aligned vector spaces, bringing similar sentences in different languages closer, and the adoption of vector space properties from the English source language

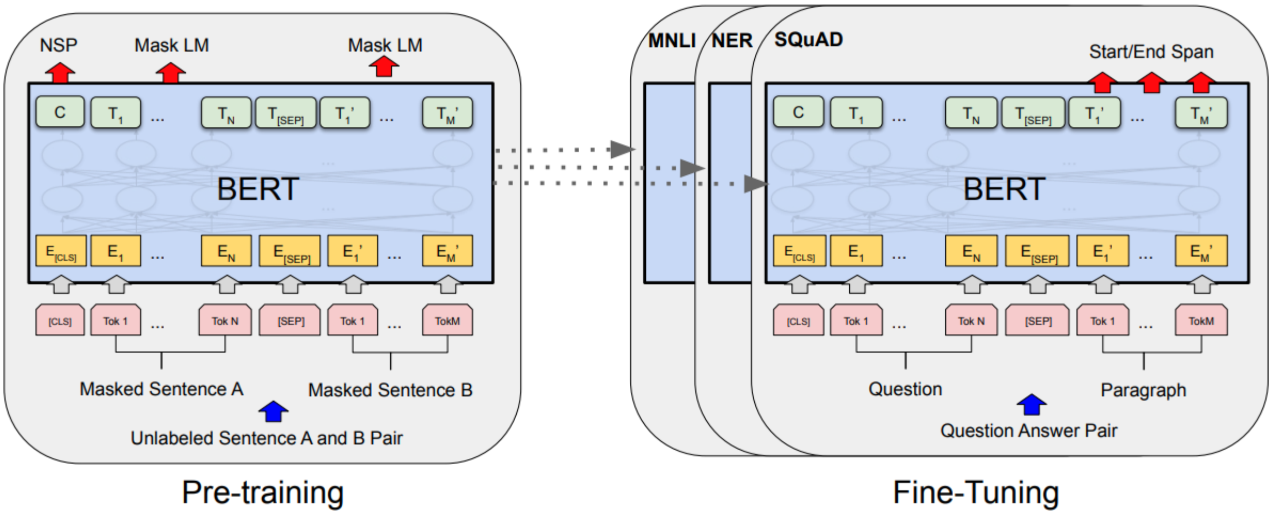


Figure 2.2: Overall pre-training and fine-tuning procedures for BERT (which are identical to BERTje) (Devlin et al., 2018)

(Reimers and Gurevych, 2020).

In the pre-training stage, a sentence transformer model is created by adding Multi-Word Expression (MWE) tokens to mBERT and training it. The model architecture uses SBERT with a regression objective function (Figure 2.5) and also includes a siamese network structure of SBERT with a classification objective function (Figure 2.4). The CoSENT method optimizes cosine similarity using contrastive learning (Reimers and Gurevych, 2019b; K. Liu, Wang, and Zhang, 2022).

In the fine-tuning stage, the sentence transformer model is created using mBERT and MWE tokens, similar to the pre-training stage. It is then trained on fine-tuned data with triplet loss and multiple negative ranking loss functions (K. Liu, Wang, and Zhang, 2022; Henderson et al., 2017). The triplet loss ensures smaller distances between correct sentence pairs compared to incorrect pairs, while the multiple negative ranking loss optimizes the model based on cosine similarity within the batch (K. Liu, Wang, and Zhang, 2022).

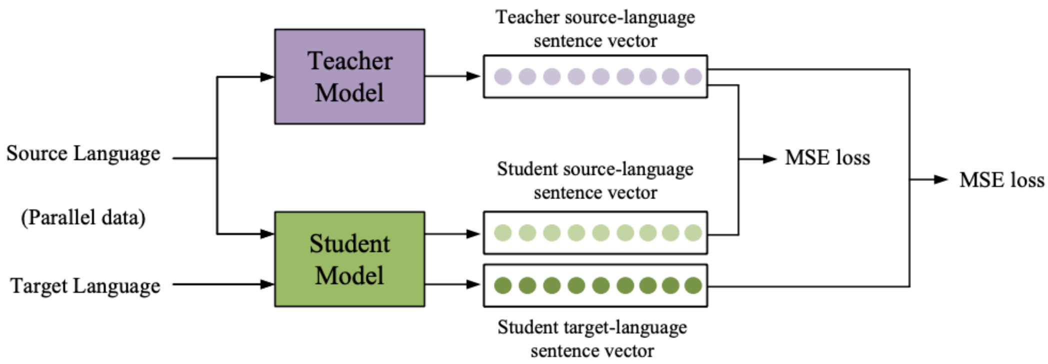


Figure 2.3: Method of making monolingual sentence embeddings multilingual using knowledge distillation (K. Liu, Wang, and Zhang, 2022)

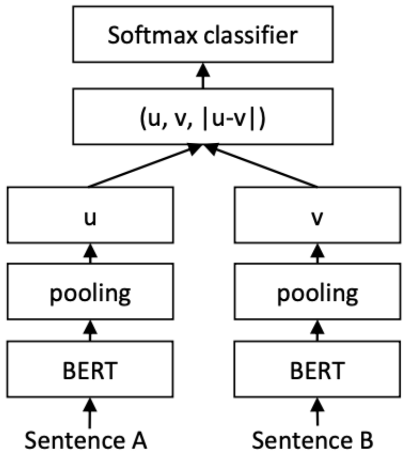


Figure 2.4: SBERT architecture with classification objective function (fine-tuning) (Reimers and Gurevych, 2019b)

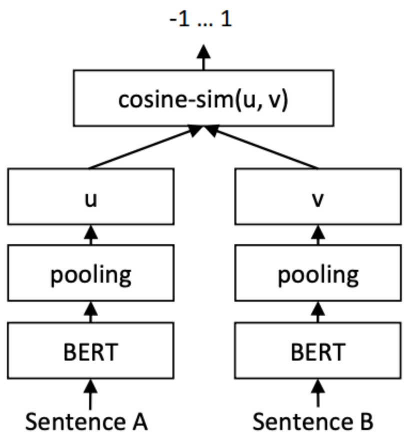


Figure 2.5: SBERT architecture for regression objective function (Reimers and Gurevych, 2019b)

3. Results

In this section, the results of the model comparison analysis are presented, which aims to determine the performance differences among the BERTje, tf-idf, and multilingual DistilUSE feature extractors. Box plots were used to compare the distribution of results and highlight the spread and consistency of each model's performance.

As depicted in Figure 3.1, the box plots demonstrate the performance of the three models across various metrics of interest, including recall at different thresholds and ATD. Notably, the BERTje model consistently outperformed the other models across all metrics, particularly excelling in the recall at the 0.9 quartiles and ATD(3.1). These preliminary findings provide a foundation for further detailed comparisons and offer valuable insights into the superiority of the BERTje model.

3.1 Psychological Dimensions

Analyzing the performance of the models on psychological dimensions, Figure 3.2a illustrates the recall curves for each dimension. We observe that the Agency_W1_ConsensusBinary, W1 redemption (both simple final and real final), and Event negvalence binary dimensions exhibit shallow recall curves. In contrast, Final Cluster A and Event posvalence binary display steeper curves, indicating the better performance of active learning compared to random screening.

Further investigation into the ATD values reveals a consistent trend of slightly lower ATD values for BERTje when compared to multilingual DistilUSE and tf-idf for all six psychological dimensions (refer to the table in Appendix A. Overall, the recall values show a trend of slightly higher values for BERTje, followed by multilingual DistilUSE and finally tf-idf.

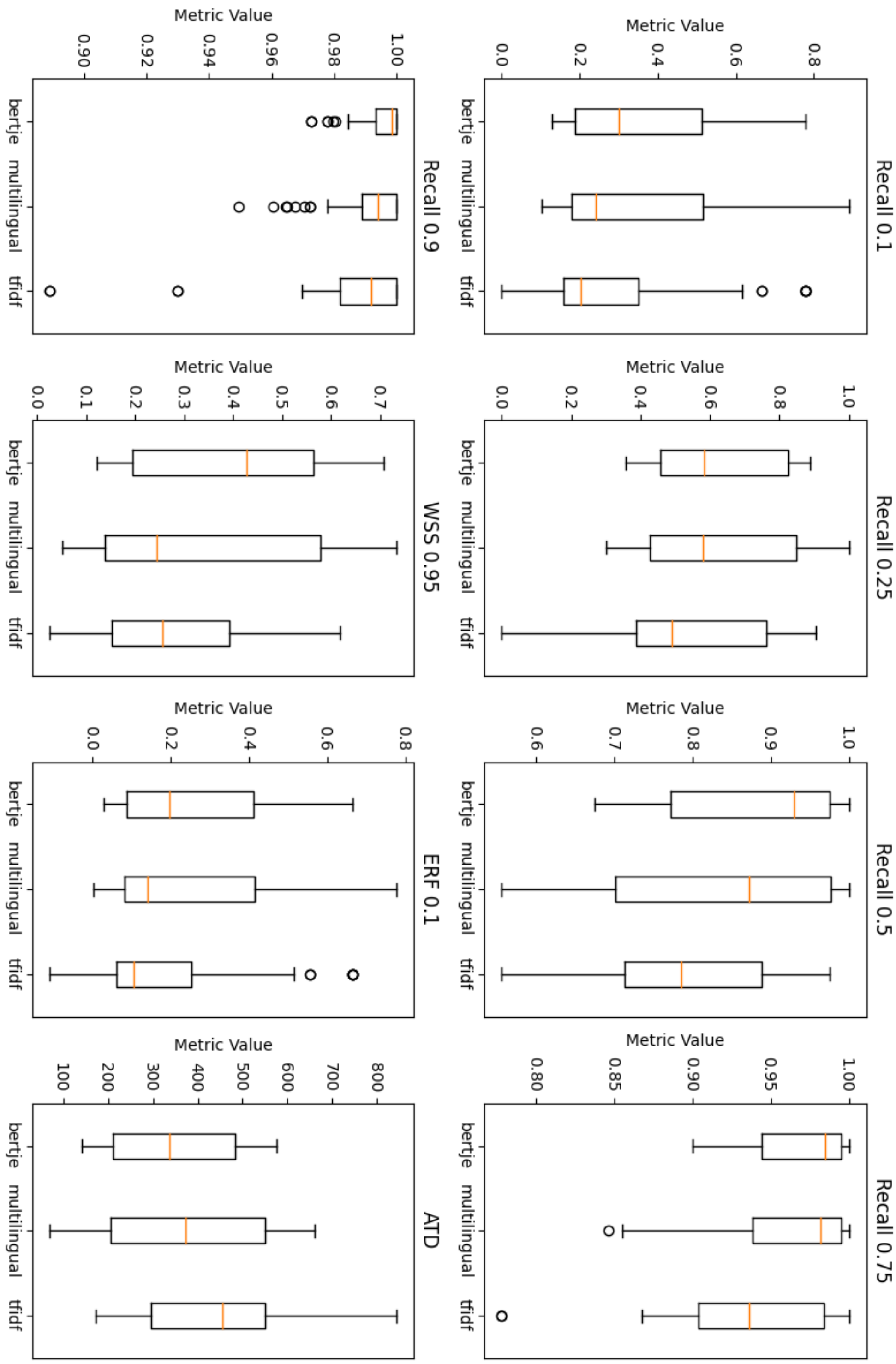


Figure 3.1: Boxplot of all the metrics for BERTje, td-idf and multilingual DistilUSE

For the Agency_W1_ConsensusBinary dimension, BERTje achieves recall values at the 0.5 quartiles ranging from 0.737 to 0.77, the multilingual extractor from 0.688 to 0.701, and tf-idf from 0.638 to 0.738. Similarly, for the Event negvalence dimension, BERTje achieves recall values at the 0.5 quartiles ranging from 0.723 to 0.749, the multilingual extractor from 0.718 to 0.749, and tf-idf from 0.67 to 0.712. Once again, BERTje outperforms the other models in this dimension. Consistently, this trend is observed across a range of labels encompassing the psychological dimensions. To obtain comprehensive recall values for each specific label, please consult the tables available in Appendix A. These additional details provide a more comprehensive understanding of the comparative performance of the models across different psychological dimensions, reinforcing the superiority of the BERTje model.

3.2 Structure dimensions

Figure 3.3a displays the recall curves for each structure dimension. Some dimensions exhibit shallow or slight recall slopes, such as the self-content religion binary and the w1N self-event connections binary. Self-content growth and self-content health recall plots show less shallow slopes but do not reach the expected steepness. On the other hand, self-content values and self-content politics display step-like recall plots, potentially due to the scarcity of 'one' data points in the dataset. Unfortunately, the self-content other recall plot is not included as the simulation resulted in a horizontal line at zero recall due to the lack of 'one' data points.

However, some structure dimensions demonstrate better and steeper recall plots. Self-content gender, self-content occupation, and self-content social exhibit quite steep recall plots compared to other dimensions. These findings indicate that active learning is more effective in capturing relevant records within these structure dimensions.

Analyzing the ATD values, we find that BERTje consistently achieves recall values ranging from 4% to 14% of the total time taken, while the multilingual extractor achieves recall values ranging from 7% to 12% of the total time taken for the simulation (refer to the table in Appendix C). This further supports the superior performance of BERTje in capturing and identifying relevant information within the structure dimensions.

BERTje consistently outperforms the multilingual and tf-idf feature extractors in the structure dimensions of the narrative identities. For instance, in the self-content gender label, BERTje achieves recall values at the 0.5 quartiles ranging from 0.975 to 0.99, whereas the multilingual extractor achieves recall values ranging from 0.98 to 0.985, and the tf-idf extractor achieves recall values ranging from 0.955 to 0.975. Similar trends of the superior performance of BERTje compared to multilingual DistilUSE and tf-idf can be observed for other labels within the structure dimension. For a comprehensive comparison of recall values across different labels, please refer to the tables provided in Appendix C.

3.3 Content dimensions

Analyzing the performance of the models on content dimensions, Figure 3.4a illustrates the recall curves for each dimension. We observe that the recall plots for event content religion and event content sex exhibit step-like slopes. Additionally, the recall plots for event content relations have a gentle slope, while the recall plots for event content other and event content health display less gentle slopes that do not resemble the steep recall plots expected from active learning. However, the recall plots for event content achievement, event content leisure, and event content school exhibit steep slopes characteristic of active learning recall plots, indicating better performance in capturing relevant records.

Analyzing the ATD values, BERTje proves to be more efficient in identifying relevant records within the content dimensions. The ATD ranges from 4% to 14% of the total time taken for the simulation, while the multilingual extractor ranges from 7% to 18% (refer the Appendix B). These results further underscore the superior performance of the BERTje model in capturing relevant information within the content dimensions.

BERTje demonstrates exceptional proficiency in capturing relevant information pertaining to narrative content dimensions, showcasing its superiority over both the multilingual and tf-idf feature extractors. For instance, when it comes to the event content achievement label, BERTje achieves recall values ranging from 0.97 to 0.976, outperforming the recall values of the multilingual extractor (0.958 to 0.97) and tf-idf (0.838 to 0.964). This pattern holds across various labels within the content dimensions. For detailed recall values for each label, please refer to the tables provided in Appendix B.

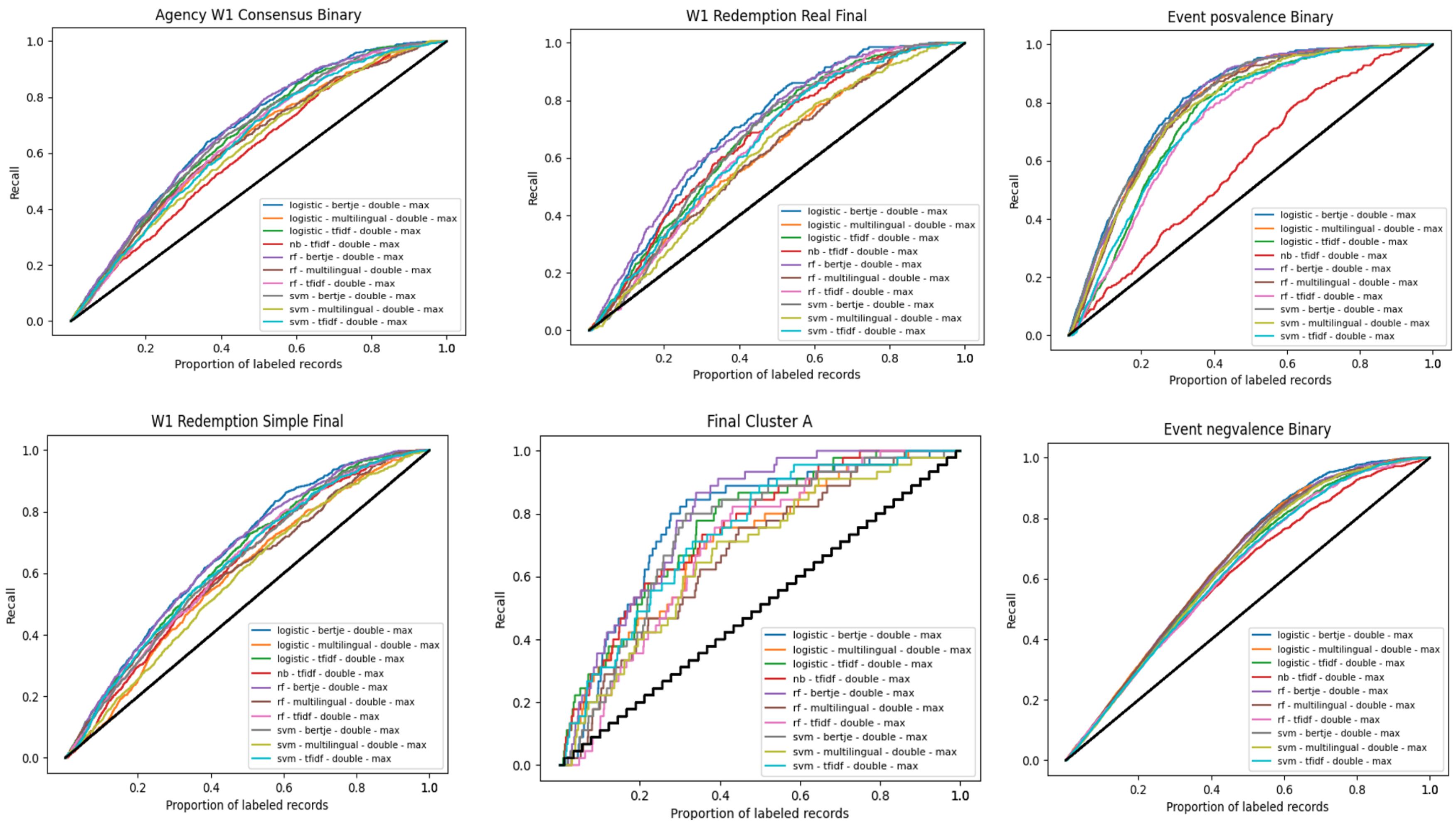


Figure 3.2: Psychological Dimensions

(a) The black line in all the recall plots represents the random sample line that is a baseline for evaluating the performance of the classifier.

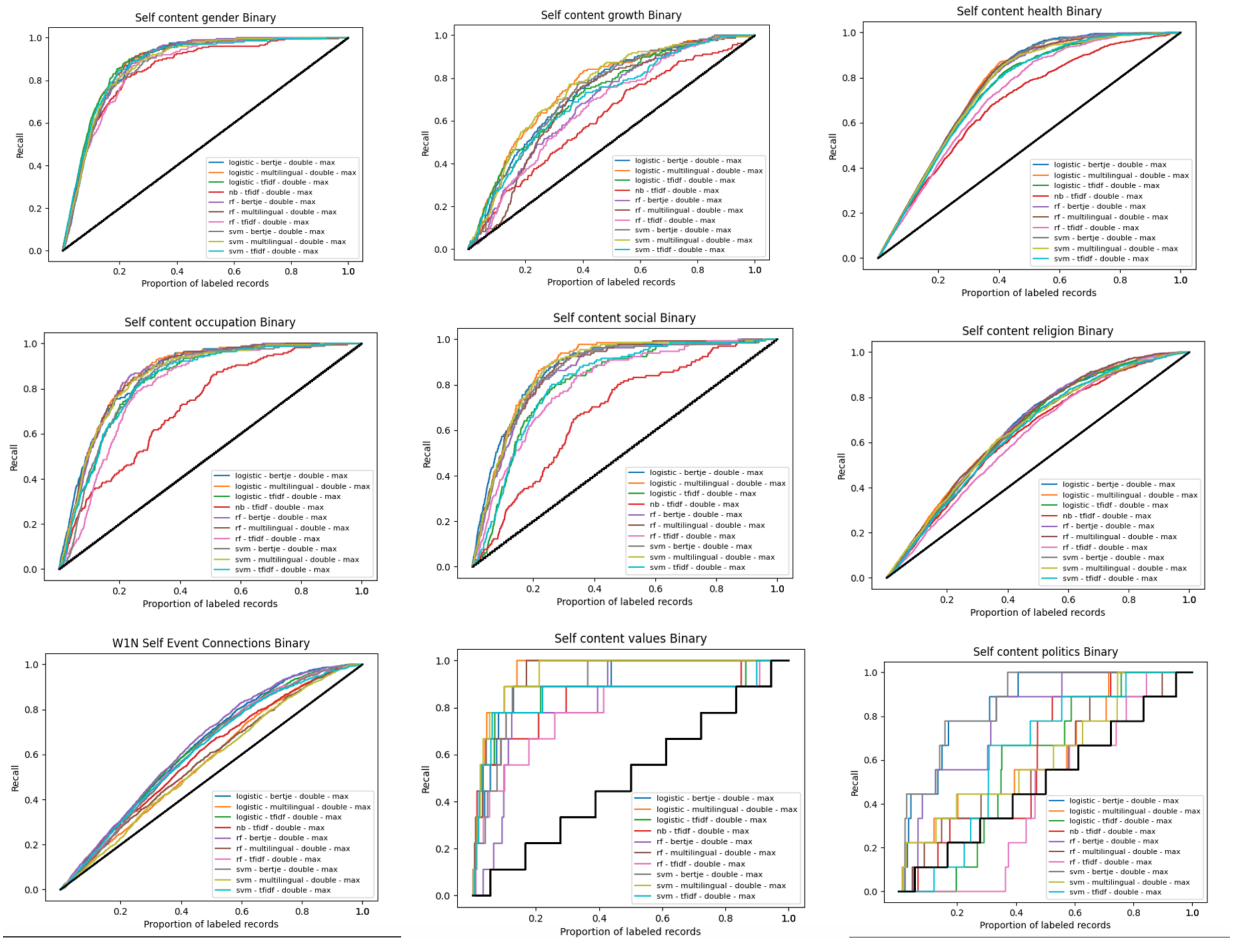
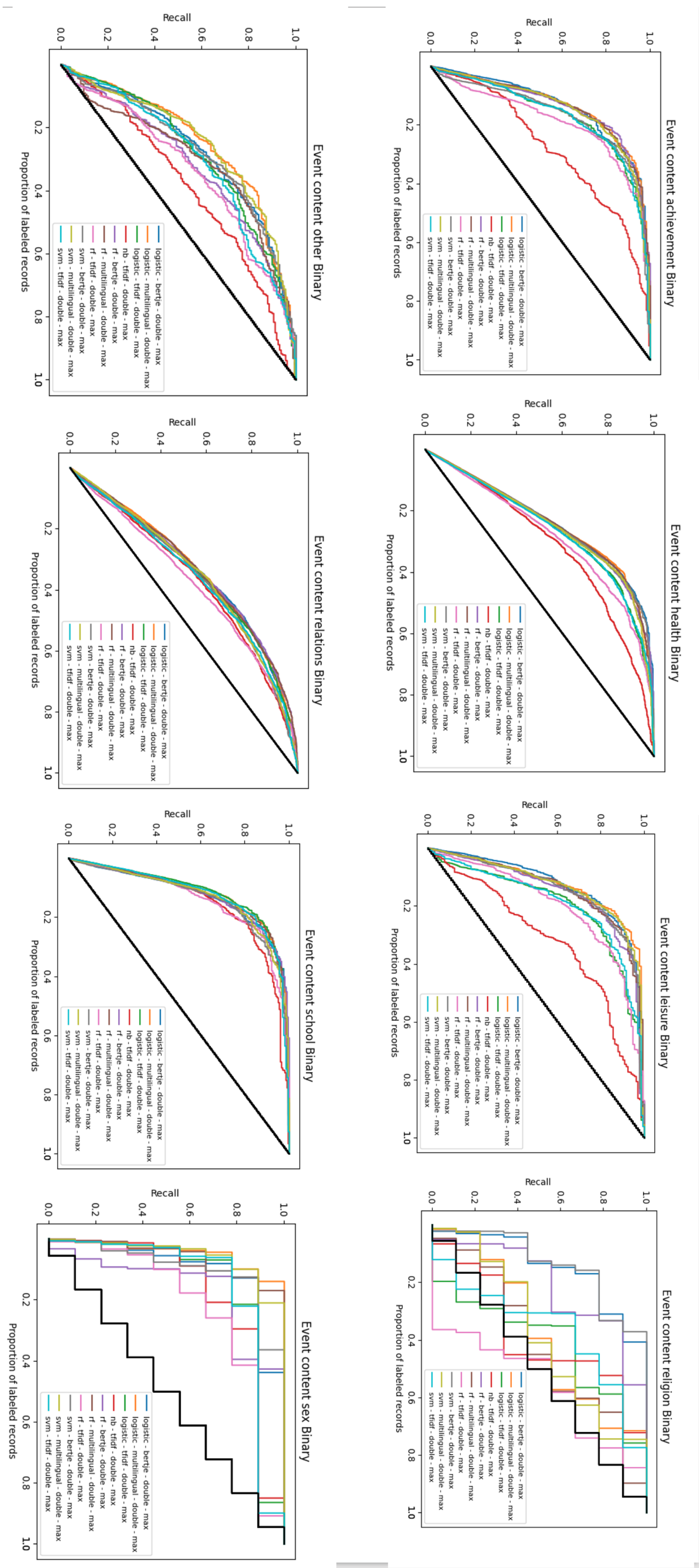


Figure 3.3: Structure Dimensions

(a) The black line in all the recall plots represents the random sample line that is a baseline for evaluating the performance of the classifier.



(a) The black line in all the recall plots represents the random sample line that is a baseline for evaluating the performance of the classifier.

Figure 3.4: Content Dimensions

4. Discussion

The present study aims to compare the effectiveness of BERTje, tf-idf, and multilingual DistilUSE feature extractors within the active learning framework for screening narratives written by Dutch adolescents. Utilizing ASReview (ASReview team, 2023) as a screening tool, this study aimed to evaluate different feature extraction methods, including monolingual, multilingual, and tf-idf approaches, to determine the most effective approach for analyzing Dutch narratives and extracting insights related to identity formation. By integrating active learning-aided tools into the systematic screening process, we aimed to uncover themes and patterns related to narrative identities in the context of the 'Project Me' initiative. As mentioned earlier in the introduction, this study builds upon the understanding that narrative identity plays a crucial role in the development of adolescents, especially for young individuals in the Netherlands facing significant decisions regarding their secondary school curriculum, which influence their sense of self (Doeselaar et al., 2020). This discussion will delve into the implications and findings of this study, shedding light on the advantages and limitations of using active learning for examining narrative identities in adolescence.

For all three dimensions, the psychological, structure and content dimensions the BERTje feature extractor outperforms the multilingual DistilUSE and tf-idf feature extractors. The analysis of the psychological dimensions revealed varying performance across different feature extractors. The BERTje feature extractor consistently achieved higher recall values for all six psychological dimensions. The findings align with previous research highlighting the effectiveness of BERT-based models in capturing semantic meaning and contextual information (De Vries et al., 2019)(Devlin et al., 2018). Moreover, the BERTje feature extractor demonstrated lower ATD values compared to the multilingual feature extractor and the default feature extractor. This indicates that BERTje is more effective in accurately identifying and capturing relevant records, resulting in reduced screening time for psychological dimensions. The superior performance of BERTje can be attributed to its advanced language modeling capabilities, specifically designed for Dutch text analysis as mentioned when discussing fine-tuning BERTje in the methodology section. By leveraging its contextual understanding and linguistic nuances, BERTje excels in extracting meaningful insights from narrative data, making it particularly well-suited for enhancing the efficiency and accuracy of narrative analysis in psychological research, specifically in the context of Dutch adolescents' narrative identities.

The analysis of the structure dimensions revealed varying slopes in the recall plots. Certain plots exhibited sub-optimal performance, characterized by step-like recall patterns, particularly for thematic elements like religion and specific aspects of self-content due to the scarcity of 'one' data points in the dataset. In contrast, dimensions related to gender, occupation, and social aspects displayed steeper slopes, indicating improved performance. The ATD values further support the effectiveness of the BERTje feature extractor in efficiently identifying relevant records within this dimension. The high recall results of the AI-aided tool for the structure dimension are useful in understanding narrative identity development. As mentioned before that self-event connections are crucial for the development of coherent narrative identity (Pasupathi, Mansour, and Brubaker, 2007). These findings contribute to our understanding of how AI-assisted screening can facilitate the exploration of narrative identities in adolescence, where thematic elements are less often mentioned in turning point narratives compared to the latter steeper sloped structure dimensions.

Regarding the context dimensions, the recall plots displayed different patterns based on the specific thematic elements. While event content related to religion and sexuality showed step-like slopes due to data scarcity, steeper slopes were observed for event content dimensions associated with achievement, leisure, and school. These findings suggest that the AI- aided tools like ASReview can effectively capture meaningful insights within the context dimensions, it shows that adolescents view the latter content dimensions as more relevant in their narratives than the former. As mentioned earlier, these dimensions represent themes of integrative meaning, emphasizing the importance of individuals connecting their past experiences to gain insight into their sense of self and contribute to psychological well-being (Kate C. McLean and Syed, 2020). Once again, the BERTje feature extractor consistently outperformed the multilingual extractor in terms of recall values for content dimensions. The ATD values further indicate the efficiency of BERTje in identifying relevant records within these dimensions.

The superior performance of BERTje in capturing the salient features within the different dimensions suggests its potential for enhancing the efficiency and accuracy of screening narratives written in Dutch. By leveraging AI-aided tools like BERTje, researchers can gain valuable insights into the complex interplay between narrative elements and identity formation.

The limitations of this paper would be a lack of data pre-processing with another limitation being the insufficient amount of data. The paper is reliant on content written by adolescents without pre-processing or cleaning which may introduce limitations in the analysis. The data may contain grammatical errors, slang, abbreviations, or other informal language patterns that can potentially affect the accuracy of natural language processing techniques used for systematic research. The dataset also contained missing values and noise in certain columns which may have affected the performance of the results. This limitation could have been mitigated by applying text-cleaning techniques or implementing a data pre-processing pipeline to enhance the data quality. Pre-processing steps such as spell-checking, normalizing abbreviations, and filtering out irrelevant content could have improved the quality of the data and enhanced the reliability of the findings. The second limitation of this study is the insufficient amount of data. The sample size of 1,580 adolescents from 'Project Me' may not be representative of the diverse population of students from various secondary school levels across the Netherlands. Consequently, this limited data may have an impact on the performance and identification of certain labels, especially those exhibiting step-like recall plots, as observed in the results section. The scarcity of '1' data points contributes to these irregular recall plots. This suggests that these labels are not relevant in turning point narratives from the perspective of adolescents. The reason for their irrelevance is attributed to the scarcity of these labels in the dataset. A larger and more representative dataset would have provided more comprehensive recall plots for all labels, with a range of steep and shallow slopes.

Future research could incorporate multi-modal features such as audio, and video alongside textual narratives to gain more comprehensive insight on narrative identity. This integration can provide additional contextual cues and enhance the model's ability to capture non-verbal expressions, visual metaphors, or emotional cues that contribute to the overall interpretation. Audio modality can enrich the understanding of the atmosphere, tone of voice, and emotions (Wu et al., 2021). Emotion nuances and state of mind can be better captured from speech patterns, intonation, or audio sentiment (Wu et al., 2021). Video modality can give insight into emotions, sincerity, and level of engagement through expressions, gestures, non-verbal cues, or body movements to give a holistic understanding of the individual's experience (Cao et al., 2014). Incorporating multi-modal features, particularly video, in narrative analysis poses challenges in navigating data privacy laws and regulations. Anonymizing voices in videos for privacy purposes may diminish the expression of emotions and tone, erasing significant cues present in speech patterns and intonation. Striking a balance between preserving privacy and capturing rich emotional cues remains a challenge for future research in multi-modal analysis.

The impact of this research is twofold. Firstly, the impact of this paper in the field of data science would be to establish BERTje as a prominent model for analyzing identity narratives in the Dutch language. Future studies may consider using BERTje as a reference point for evaluating the performance of subsequent models in similar tasks, considering its demonstrated effectiveness as a strong candidate for comparison. This means that if you have mono-language data, a mono-language feature extractor should be used for NLP tasks. Secondly, this paper is relevant to practitioners, psychologists, and researchers working with narratives, as it demonstrates the superior performance of the BERTje feature extractor. The integration of AI-aided screening methods, such as ASReview, with developmental psychology practices accelerates the process of reviewing narratives and facilitates informed decisions, tailored support strategies, and further exploration in the field, thereby enhancing our understanding of the formation of identity. Adopting software originally developed for systematic reviewing by practitioners signifies the potential for interdisciplinary collaboration and allows for the efficient extraction of relevant features and valuable information. This enhances understanding of narrative identities and their application in psychological interventions.

5. Conclusion

In conclusion, this study demonstrates the effectiveness of BERTje, tf-idf, and distilUSE multilingual feature extractors within the active learning framework for analyzing the narrative identities of Dutch adolescents. BERTje consistently outperforms other extractors across psychological, structural, and content dimensions in Dutch language text, capturing semantic meaning and contextual information to extract insights on identity formation. Its higher recall values and lower ATD indicate superior performance in capturing relevant records and reducing screening time.

The study suggests exploring the incorporation of multi-modal features, such as audio and video, to enhance the understanding of narrative identity, considering privacy concerns associated with video modality. Furthermore, this research establishes BERTje as a prominent model for analyzing narrative identities in Dutch adolescents and highlights the potential of AI-aided tools like ASReview for efficient feature extraction in psychological research. Future studies should explore integrating multi-modal features while addressing challenges related to data privacy laws. These findings contribute to advancing language models, facilitating informed decision-making, and enhancing our understanding of narrative identities in psychological interventions.

5.1 Acknowledgement

I am sincerely grateful to my supervisors, Rens van de Schoot and Laura Hofstee, for their exceptional guidance and unwavering support throughout this thesis. Their expertise and dedication have been indispensable to the completion of this research. I would also like to express my gratitude to researcher Jaap Denissen for his invaluable inspiration. Their support has been instrumental in the successful completion of this thesis. Additionally, I extend my heartfelt appreciation to Utrecht University for providing the necessary resources.

Bibliography

Laird, Robert D. (2013). "Adolescence". In: *Oxford Bibliographies Online*. DOI: 10.1093/obo/9780199828340-0008.

Erikson, Erik H (1950). *Childhood and society*. WW Norton & Company.

Klimstra, Theo (2013). "Adolescent personality development and identity formation". In: *Child Development Perspectives* 7.2, pp. 80–84.

Erikson, E. H. (1968). *Identity: Youth and Crisis*. New York, NY: Norton.

Doeselaar, Lotte van et al. (2020). "Adolescents' identity formation: Linking the narrative and the dual-cycle approach". In: *Journal of youth and adolescence* 49, pp. 818–835.

Tilburg University (2023). *Tilburg University - Department of Developmental Psychology - Research: Project I&K - Research in Practice*. URL: https://www.tilburguniversity.edu/nl/over/schools/socialsciences/organisatie/departementen/ontwikkelingspsychologie/onderzoek/project-ik/onderzoek-de-praktijk.

ASReview team (2023). *ASReview: Active Learning for Systematic Reviews*. URL: https://asreview.nl/.

Van den Brand, S. A. G. E. and R. van de Schoot (Sept. 2021). *A Systematic Review on Studies Evaluating the Performance of Active Learning Compared to Human Reading for Systematic Review Data*. URL: https://doi.org/10.17605/OSF.IO/T9HGM.

Singer, Jefferson A (2004). "Narrative identity and meaning making across the adult lifespan: An introduction". In: *Journal of personality* 72.3, pp. 437–460.

McAdams, D. P. (1993). *The Stories We Live By Personal Myths and the Making of the Self*. New York, NY: Guilford Press.

McAdams, Dan P and Kate C McLean (2013). "Narrative identity". In: *Current directions in psychological science* 22.3, pp. 233–238.

Adler, Jonathan M et al. (2017). "Research methods for studying narrative identity: A primer". In: *Social Psychological and Personality Science* 8.5, pp. 519–527.

McLean, Kate C et al. (2020). "The empirical structure of narrative identity: The initial Big Three." In: *Journal of personality and social psychology* 119.4, p. 920.

McLean, Kate C. and Moin Syed (2020). "Narrative Identity". In: *The Encyclopedia of Child and Adolescent Development*. John Wiley & Sons, Ltd. ISBN: 9781119171492. DOI: https://doi.org/10.1002/9781119171492.wecad454.

Habermas, Tilmann and Elaine Reese (2015). "Getting a life takes time: The development of the life story in adolescence, its precursors and consequences". In: *Human Development* 58.3, pp. 172–201.

Pasupathi, Monisha, Emma Mansour, and Jed R Brubaker (2007). "Developing a life story: Constructing relations between self and experience in autobiographical narratives". In: *Human development* 50.2-3, pp. 85–110.

De Vries, Wietse et al. (2019). "Bertje: A dutch bert model". In: *arXiv preprint arXiv:1912.09582*.

Van Doeselaar, Lotte (Aug. 2019). *Adolescents' Identity Formation: Linking the Narrative and the Dual-Cycle Approach*. URL: https://osf.io/tnyaf.

ASReview developers (2023). *ASReview Documentation*. URL: https://asreview.readthedocs.io/en/latest/about.html.

Teijema, Jelle et al. (2023). *ASReview Makita: a workflow generator for simulation studies using the command line interface of ASReview LAB*. URL: https://github.com/asreview/asreview-makita.

developers, ASReview LAB (n.d.). *ASReview Insights - Insights and plotting tool for the ASReview project*. URL: https://github.com/asreview/asreview-insights.

Vries, Wietse de et al. (Dec. 2019). *BERTje: A Dutch BERT Model*. arXiv:1912.09582. URL: http://arxiv.org/abs/1912.09582.

Reimers, Nils and Iryna Gurevych (Nov. 2019a). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: http://arxiv.org/abs/1908.10084.

Teijema, J. (2021). *BERTje_fe*. https://github.com/jteijema/BERTje_fe.

– (2023). *Multilingual-Feature-Extractor*. https://github.com/jteijema/multilingual-feature-extractor.

Atwani, Kajol (2023). *feature-extractors*. https://github.com/KajolAtwani/feature-extractors.

Ferdinands, Gerbrich et al. (2020). *Active learning for screening prioritization in systematic reviews - A simulation study*. DOI: 10.31219/osf.io/w6qbg. URL: osf.io/w6qbg.

Stecanella, Bruno (May 2019). *Understanding TF-ID: A simple introduction*. URL: https://monkeylearn.com/blog/what-is-tf-idf/.

Liu, Qing et al. (2018). "Text features extraction based on TF-IDF associating semantic". In: *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*. IEEE, pp. 2338–2343.

Analytics Vidhya (2021). *How sklearn's Tfidfvectorizer Calculates tf-idf Values*. URL: https://www.analyticsvidhya.com/blog/2021/11/how-sklearns-tfidfvectorizer-calculates-tf-idf-values/.

Devlin, Jacob et al. (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805*.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Oostdijk, Nelleke et al. (2013). "The construction of a 500-million-word reference corpus of contemporary written Dutch". In: *Essential speech and language technology for Dutch: Results by the STEVIN programme*, pp. 219–247.

Ordelman, Roeland et al. (2007). "Twnc: a multifaceted Dutch news corpus". In: *ELRA Newsletter* 12.3/4, pp. 4–7.

Sang, Erik F and Fien De Meulder (2003). "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". In: *arXiv preprint cs/0306050*.

Van Noord, Gertjan et al. (2013). "Large scale syntactic annotation of written Dutch: Lassy". In: *Essential speech and language technology for Dutch: results by the STEVIN programme*, pp. 147–164.

Delaere, Isabelle, Veronique Hoste, and Paola Monachesi (2009). "Cultivating trees: Adding several semantic layers to the Lassy treebank in SoNaR". In: *7th International workshop on Treebanks and Linguistic Theories (TLT-7)*. LOT (Landelijke Onderzoekschool Taalwetenschap), pp. 135–146.

Tiedemann, Jörg (2012). "Parallel data, tools and interfaces in OPUS." In: *Lrec*. Vol. 2012. Citeseer, pp. 2214–2218.

Koehn, Philipp (2005). "Europarl: A parallel corpus for statistical machine translation". In: *Proceedings of machine translation summit x: papers*, pp. 79–86.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen (2016). "The united nations parallel corpus v1. 0". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534.

Agic, Željko and Ivan Vulic (2019). "JW300: A wide-coverage parallel corpus for low-resource languages". In: Association for Computational Linguistics.

Schwenk, Holger et al. (2019). "Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia". In: *arXiv preprint arXiv:1907.05791*.

Reimers, Nils and Iryna Gurevych (Apr. 2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". In: *arXiv preprint arXiv:2004.09813*. URL: `http://arxiv.org/abs/2004.09813`.

Yang, Yinfei et al. (2019). "Multilingual universal sentence encoder for semantic retrieval". In: *arXiv preprint arXiv:1907.04307*.

Liu, Kuanghong, Jin Wang, and Xuejie Zhang (2022). "Ynu-hpcc at semeval-2022 task 2: Representing multilingual idiomaticity based on contrastive learning". In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 211–216.

Reimers, Nils and Iryna Gurevych (2019b). "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084*.

Henderson, Matthew et al. (2017). "Efficient natural language response suggestion for smart reply". In: *arXiv preprint arXiv:1705.00652*.

Wu, Jingyao et al. (2021). "Multimodal affect models: An investigation of relative salience of audio and visual cues for emotion prediction". In: *Frontiers in Computer Science* 3, p. 767767.

Cao, Houwei et al. (2014). "Crema-d: Crowd-sourced emotional multimodal actors dataset". In: *IEEE transactions on affective computing* 5.4, pp. 377–390.

Appendices

A Psychological Dimension

Agency W1 Consensus	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
svm_bertje	0.174	0.455	0.737	0.926	0.987	0.138	0.076	510.737
svm_tfidf	0.171	0.401	0.716	0.922	0.978	0.141	0.072	541.506
rf_tfidf	0.158	0.414	0.722	0.929	0.981	0.161	0.059	533.944
logistic_multilingual	0.178	0.421	0.701	0.892	0.972	0.088	0.08	549.015
rf_multilingual	0.178	0.427	0.688	0.889	0.965	0.081	0.08	552.831
logistic_tfidf	0.189	0.434	0.738	0.928	0.985	0.174	0.091	506.972
nb_tfidf	0.167	0.347	0.638	0.889	0.98	0.103	0.069	597.043
svm_multilingual	0.167	0.39	0.668	0.887	0.972	0.12	0.069	573.571
logistic_bertje	0.186	0.456	0.764	0.95	0.993	0.203	0.087	483.963
rf_bertje	0.195	0.456	0.77	0.939	0.991	0.176	0.096	483.128

Table 1: Agency W1 Consensus Metrics

Event Negvalence	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_tfidf	0.151	0.378	0.712	0.926	0.987	0.153	0.053	546.912
logistic_bertje	0.153	0.377	0.749	0.959	0.993	0.231	0.055	521.109
svm_bertje	0.147	0.377	0.723	0.938	0.984	0.165	0.048	541.094
nb_tfidf	0.152	0.368	0.67	0.882	0.97	0.098	0.054	584.106
rf_bertje	0.151	0.392	0.741	0.944	0.989	0.191	0.053	526.535
rf_multilingual	0.153	0.39	0.749	0.938	0.989	0.183	0.055	525.482
svm_tfidf	0.146	0.368	0.698	0.92	0.984	0.141	0.047	561.563
logistic_multilingual	0.155	0.391	0.741	0.945	0.992	0.187	0.056	523.027
rf_tfidf	0.15	0.368	0.69	0.917	0.978	0.146	0.052	565.353
svm_multilingual	0.143	0.372	0.718	0.943	0.984	0.18	0.045	544.28

Table 2: Event Negvalence Metrics

Event Posvalence	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
svm_bertje	0.347	0.694	0.939	0.983	0.997	0.43	0.249	304.841
svm_multilingual	0.309	0.673	0.896	0.983	0.994	0.363	0.211	332.329
svm_tfidf	0.243	0.581	0.89	0.98	0.991	0.313	0.145	382.422
logistic_tfidf	0.194	0.598	0.899	0.983	0.994	0.316	0.095	374.786
logistic_bertje	0.364	0.723	0.939	0.986	0.994	0.437	0.266	292.61
nb_tfidf	0.147	0.335	0.627	0.887	0.974	0.107	0.049	614.156
logistic_multilingual	0.315	0.694	0.934	0.983	0.994	0.425	0.217	313.223
rf_multilingual	0.298	0.676	0.925	0.986	0.994	0.371	0.199	324.017
rf_tfidf	0.199	0.555	0.867	0.983	0.991	0.314	0.101	399.723
rf_bertje	0.332	0.679	0.931	0.986	0.994	0.419	0.234	306.185

Table 3: Event Posvalence Metrics

Final Cluster A	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_multilingual	0.311	0.489	0.778	0.956	1	0.226	0.222	438.489
logistic_tfidf	0.311	0.6	0.867	0.978	1	0.29	0.222	367.467
logistic_bertje	0.267	0.711	0.889	0.956	0.978	0.304	0.178	337.044
nb_tfidf	0.289	0.6	0.844	1	1	0.28	0.2	379.644
rf_bertje	0.356	0.578	0.933	1	1	0.46	0.267	310.422
svm_multilingual	0.222	0.422	0.756	0.911	0.978	0.131	0.133	495.511
svm_tfidf	0.289	0.556	0.889	0.956	1	0.346	0.2	393.622
svm_bertje	0.2	0.622	0.844	0.956	0.978	0.282	0.111	394.4
rf_multilingual	0.178	0.467	0.756	0.956	1	0.196	0.089	496.356
rf_tfidf	0.111	0.467	0.822	0.956	1	0.284	0.022	459.556

Table 4: Final Cluster A Metrics

W1 Redemption Real Final	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_tfidf	0.165	0.408	0.768	0.949	0.989	0.201	0.066	497.574
rf_bertje	0.217	0.529	0.79	0.967	0.985	0.258	0.118	447.478
svm_multilingual	0.132	0.335	0.688	0.886	0.978	0.09	0.033	592.555
nb_tfidf	0.169	0.452	0.743	0.934	0.989	0.181	0.07	507.938
rf_multilingual	0.143	0.371	0.658	0.89	0.993	0.146	0.044	584.309
rf_tfidf	0.118	0.382	0.746	0.963	0.989	0.212	0.018	530.368
svm_tfidf	0.176	0.386	0.746	0.93	0.985	0.158	0.077	527.537
logistic_multilingual	0.121	0.375	0.654	0.89	0.989	0.139	0.022	579.283
svm_bertje	0.129	0.426	0.787	0.934	0.996	0.165	0.029	511.327
logistic_bertje	0.188	0.496	0.82	0.985	0.996	0.255	0.088	450.313

Table 5: W1 Redemption Real Final Metrics

W1 Redemption Simple Final	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
rf_multilingual	0.174	0.394	0.638	0.871	0.98	0.099	0.075	590.299
logistic_multilingual	0.101	0.358	0.644	0.861	0.968	0.076	0.002	615.719
svm_tfidf	0.184	0.41	0.685	0.905	0.98	0.118	0.085	551.881
nb_tfidf	0.143	0.364	0.675	0.899	0.978	0.106	0.044	578.705
rf_bertje	0.194	0.438	0.733	0.931	0.992	0.181	0.095	512.376
logistic_tfidf	0.192	0.412	0.713	0.917	0.984	0.153	0.093	536.125
logistic_bertje	0.18	0.434	0.725	0.939	0.994	0.19	0.081	507.212
svm_multilingual	0.121	0.317	0.624	0.865	0.949	0.05	0.022	631.788
rf_tfidf	0.172	0.386	0.695	0.907	0.974	0.117	0.073	560.937
svm_bertje	0.166	0.384	0.675	0.913	0.98	0.137	0.067	563.828

Table 6: W1 Redemption Simple Final Metrics

B Content Dimension

Event_content_achievements	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
rf_multilingual_0	0.515	0.850	0.964	1.000	1.000	0.561	0.413	209.299
rf_tfidf_0	0.281	0.772	0.964	0.988	1.000	0.502	0.180	289.766
svm_bertje_0	0.431	0.796	0.970	0.988	0.994	0.526	0.329	253.353
rf_bertje_0	0.509	0.868	0.970	1.000	1.000	0.553	0.407	209.874
svm_tfidf_0	0.395	0.796	0.952	0.988	1.000	0.464	0.293	253.311
logistic_bertje_0	0.533	0.808	0.976	0.994	1.000	0.566	0.431	208.240
nb_tfidf_0	0.317	0.491	0.838	0.958	0.994	0.256	0.216	419.976
logistic_tfidf_0	0.413	0.790	0.958	0.994	1.000	0.483	0.311	249.515
logistic_multilingual_0	0.533	0.838	0.970	1.000	1.000	0.584	0.431	204.784
svm_multilingual_0	0.515	0.844	0.958	0.994	1.000	0.517	0.413	219.323

Table 7: Event Content Achievements Metrics

Event_content_health	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
logistic_bertje_0	0.237	0.582	0.934	0.995	0.998	0.429	0.138	353.5
nb_tfidf_0	0.225	0.487	0.785	0.939	0.977	0.171	0.127	469.845
svm_multilingual_0	0.237	0.571	0.911	0.979	0.995	0.347	0.138	373.168
rf_bertje_0	0.234	0.571	0.911	0.992	0.998	0.383	0.135	365.238
rf_tfidf_0	0.207	0.516	0.855	0.972	0.998	0.265	0.109	423.12
svm_tfidf_0	0.229	0.554	0.878	0.984	0.993	0.323	0.13	391.531
svm_bertje_0	0.235	0.571	0.929	0.987	0.998	0.417	0.137	361.916
rf_multilingual_0	0.24	0.581	0.924	0.982	0.995	0.38	0.141	360.934
logistic_tfidf_0	0.234	0.563	0.878	0.975	0.992	0.3	0.135	390.584
logistic_multilingual_0	0.24	0.579	0.923	0.987	0.995	0.371	0.141	359.201

Table 8: Event Content Health Metrics

Event_content_leisure	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
rf_bertje_0	0.473	0.824	0.962	0.977	1	0.563	0.374	226.458
nb_tfidf_0	0.191	0.443	0.817	0.916	0.985	0.163	0.092	480.45
rf_multilingual_0	0.511	0.847	0.977	0.992	0.992	0.544	0.412	207.42
logistic_multilingual_0	0.489	0.878	0.985	0.992	0.992	0.624	0.389	189.71
svm_multilingual_0	0.504	0.885	0.985	0.985	0.992	0.597	0.405	198.328
rf_tfidf_0	0.389	0.718	0.908	0.985	1	0.328	0.29	300.366
svm_bertje_0	0.511	0.84	0.969	0.977	1	0.575	0.412	215.756
logistic_tfidf_0	0.351	0.763	0.916	0.985	0.985	0.367	0.252	294.817
logistic_bertje_0	0.573	0.863	0.977	0.985	1	0.571	0.473	191.863
svm_tfidf_0	0.321	0.779	0.924	0.985	0.985	0.392	0.221	285.389

Table 9: Event Content Leisure Metrics

Event_content_other	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
rf_bertje_0	0.191	0.49	0.777	0.936	0.994	0.188	0.089	479.236
svm_bertje_0	0.274	0.561	0.834	0.955	1	0.235	0.172	406.223
logistic_bertje_0	0.299	0.573	0.854	0.955	1	0.213	0.197	401.49
nb_tfidf_0	0.172	0.389	0.669	0.879	0.93	0.025	0.07	579.981
svm_multilingual_0	0.261	0.637	0.873	0.955	0.994	0.244	0.159	375.758
logistic_multilingual_0	0.312	0.624	0.866	0.962	0.994	0.232	0.21	368.548
rf_multilingual_0	0.115	0.516	0.834	0.962	0.987	0.207	0.013	455.083
logistic_tfidf_0	0.312	0.541	0.783	0.949	0.987	0.2	0.21	420.452
svm_tfidf_0	0.274	0.548	0.758	0.943	0.994	0.183	0.172	449.949
rf_tfidf_0	0.172	0.452	0.758	0.936	0.994	0.177	0.07	506.382

Table 10: Event Content Other Metrics

Event_content_relations	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_bertje_0	0.18	0.431	0.772	0.932	0.985	0.163	0.08	502.222
rf_bertje_0	0.174	0.435	0.764	0.944	0.994	0.195	0.074	496.059
logistic_multilingual_0	0.186	0.45	0.755	0.938	0.989	0.168	0.086	496.295
rf_tfidf_0	0.139	0.373	0.693	0.906	0.982	0.135	0.039	565.746
svm_tfidf_0	0.166	0.4	0.737	0.923	0.973	0.146	0.066	531.065
rf_multilingual_0	0.19	0.452	0.76	0.947	0.992	0.195	0.091	495.116
svm_multilingual_0	0.181	0.429	0.733	0.917	0.979	0.129	0.082	519.408
logistic_tfidf_0	0.162	0.429	0.749	0.929	0.983	0.161	0.062	513.707
nb_tfidf_0	0.16	0.4	0.711	0.903	0.973	0.108	0.06	549.523
svm_bertje_0	0.162	0.42	0.74	0.909	0.973	0.122	0.062	529.595

Table 11: Event Content Relations Metrics

Event_content_religion	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_tfidf_0	0	0.111	0.667	0.889	1	0.246	-0.111	619.778
logistic_multilingual_0	0.222	0.444	0.556	1	1	0.127	0.111	561.556
logistic_bertje_0	0.444	0.778	1	1	1	0.524	0.333	217.444
svm_tfidf_0	0	0.333	0.778	0.889	1	0.278	-0.111	549.889
rf_tfidf_0	0	0	0.556	0.778	1	0.058	-0.111	843.556
svm_multilingual_0	0.222	0.444	0.556	1	1	0.09	0.111	571.444
svm_bertje_0	0.444	0.778	1	1	1	0.501	0.333	204.444
rf_bertje_0	0.444	0.556	0.889	1	1	0.5	0.333	317.222
nb_tfidf_0	0.111	0.333	0.778	1	1	0.31	0	583.222
rf_multilingual_0	0.222	0.333	0.556	0.889	1	0.183	0.111	626

Table 12: Event Content Religion Metrics

Event_content_school	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
rf_multilingual_0	0.555	0.905	0.98	0.995	0.995	0.599	0.455	182.785
svm_tfidf_0	0.59	0.89	0.97	0.995	0.995	0.604	0.49	182.67
svm_multilingual_0	0.56	0.855	0.985	0.995	1	0.577	0.46	190.82
logistic_tfidf_0	0.615	0.905	0.97	0.995	0.995	0.607	0.515	172.62
nb_tfidf_0	0.54	0.825	0.955	0.985	0.995	0.476	0.44	223.67
logistic_bertje_0	0.555	0.89	0.985	0.995	1	0.617	0.455	188.105
svm_bertje_0	0.525	0.835	0.975	0.995	1	0.581	0.425	199.955
logistic_multilingual_0	0.58	0.905	0.985	0.995	1	0.595	0.48	175.395
rf_tfidf_0	0.51	0.87	0.975	1	1	0.511	0.41	210.6
rf_bertje_0	0.545	0.88	0.99	1	1	0.617	0.445	184.085

Table 13: Event Content School Metrics

Event_content_sex	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_bertje_0	0.778	0.889	1	1	1	0.705	0.667	141.222
logistic_tfidf_0	0.778	0.889	0.889	0.889	1	0.619	0.667	217.222
nb_tfidf_0	0.667	0.778	0.889	0.889	1	0.538	0.556	251.556
svm_bertje_0	0.667	0.889	1	1	1	0.708	0.556	144.556
rf_bertje_0	0.556	0.778	1	1	1	0.439	0.444	242.333
rf_multilingual_0	0.667	1	1	1	1	0.707	0.556	88
svm_multilingual_0	0.778	1	1	1	1	0.734	0.667	79.222
svm_tfidf_0	0.778	0.889	0.889	0.889	1	0.613	0.667	221.667
logistic_multilingual_0	0.889	1	1	1	1	0.734	0.778	69.111
rf_tfidf_0	0.444	0.667	0.889	0.889	0.889	0.419	0.333	329.778

Table 14: Event Content Sex Metrics

C Structure Dimension

Self_content_gender	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
logistic_multilingual_0	0.58	0.905	0.985	0.995	1	0.595	0.48	175.395
logistic_bertje_0	0.555	0.89	0.985	0.995	1	0.617	0.455	188.105
svm_tfidf_0	0.59	0.89	0.97	0.995	0.995	0.604	0.49	182.67
svm_bertje_0	0.525	0.835	0.975	0.995	1	0.581	0.425	199.955
logistic_tfidf_0	0.615	0.905	0.97	0.995	0.995	0.607	0.515	172.62
nb_tfidf_0	0.54	0.825	0.955	0.985	0.995	0.476	0.44	223.67
svm_multilingual_0	0.56	0.855	0.985	0.995	1	0.577	0.46	190.82
rf_multilingual_0	0.555	0.905	0.98	0.995	0.995	0.599	0.455	182.785
rf_tfidf_0	0.51	0.87	0.975	1	1	0.511	0.41	210.6
rf_bertje_0	0.545	0.88	0.99	1	1	0.617	0.445	184.085

Table 15: Self Content Gender Metrics

Self_content_growth	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
svm_multilingual_0	0.261	0.637	0.873	0.955	0.994	0.244	0.159	375.758
rf_tfidf_0	0.172	0.452	0.758	0.936	0.994	0.177	0.07	506.382
svm_tfidf_0	0.274	0.548	0.758	0.943	0.994	0.183	0.172	449.949
logistic_bertje_0	0.299	0.573	0.854	0.955	1	0.213	0.197	401.49
svm_bertje_0	0.274	0.561	0.834	0.955	1	0.235	0.172	406.223
nb_tfidf_0	0.172	0.389	0.669	0.879	0.93	0.025	0.07	579.981
rf_multilingual_0	0.115	0.516	0.834	0.962	0.987	0.207	0.013	455.083
logistic_tfidf_0	0.312	0.541	0.783	0.949	0.987	0.2	0.21	420.452
logistic_multilingual_0	0.312	0.624	0.866	0.962	0.994	0.232	0.21	368.548
rf_bertje_0	0.191	0.49	0.777	0.936	0.994	0.188	0.089	479.236

Table 16: Self Content Growth Metrics

Self_content_health	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
svm_tfidf_0	0.229	0.554	0.878	0.984	0.993	0.323	0.13	391.531
svm_multilingual_0	0.237	0.571	0.911	0.979	0.995	0.347	0.138	373.168
rf_multilingual_0	0.24	0.581	0.924	0.982	0.995	0.38	0.141	360.934
svm_bertje_0	0.235	0.571	0.929	0.987	0.998	0.417	0.137	361.916
rf_tfidf_0	0.207	0.516	0.855	0.972	0.998	0.265	0.109	423.12
logistic_multilingual_0	0.24	0.579	0.923	0.987	0.995	0.371	0.141	359.201
logistic_tfidf_0	0.234	0.563	0.878	0.975	0.992	0.3	0.135	390.584
rf_bertje_0	0.234	0.571	0.911	0.992	0.998	0.383	0.135	365.238
nb_tfidf_0	0.225	0.487	0.785	0.939	0.977	0.171	0.127	469.845
logistic_bertje_0	0.237	0.582	0.934	0.995	0.998	0.429	0.138	353.5

Table 17: Self Content Health Metrics

Self_content_occupation	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
rf_multilingual_0	0.515	0.85	0.964	1	1	0.561	0.413	209.299
rf_tfidf_0	0.281	0.772	0.964	0.988	1	0.502	0.18	289.766
logistic_tfidf_0	0.413	0.79	0.958	0.994	1	0.483	0.311	249.515
svm_multilingual_0	0.515	0.844	0.958	0.994	1	0.517	0.413	219.323
svm_bertje_0	0.431	0.796	0.97	0.988	0.994	0.526	0.329	253.353
rf_bertje_0	0.509	0.868	0.97	1	1	0.553	0.407	209.874
svm_tfidf_0	0.395	0.796	0.952	0.988	1	0.464	0.293	253.311
logistic_bertje_0	0.533	0.808	0.976	0.994	1	0.566	0.431	208.24
nb_tfidf_0	0.317	0.491	0.838	0.958	0.994	0.256	0.216	419.976
logistic_multilingual_0	0.533	0.838	0.97	1	1	0.584	0.431	204.784

Table 18: Self Content Occupation Metrics

Self_content_politics	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
rf_multilingual_0	0.222	0.333	0.556	0.889	1	0.183	0.111	626
logistic_multilingual_0	0.222	0.444	0.556	1	1	0.127	0.111	561.556
svm_multilingual_0	0.222	0.444	0.556	1	1	0.09	0.111	571.444
rf_bertje_0	0.444	0.556	0.889	1	1	0.5	0.333	317.222
rf_tfidf_0	0	0	0.556	0.778	1	0.058	-0.111	843.556
logistic_bertje_0	0.444	0.778	1	1	1	0.524	0.333	217.444
nb_tfidf_0	0.111	0.333	0.778	1	1	0.31	0	583.222
svm_bertje_0	0.444	0.778	1	1	1	0.501	0.333	204.444
logistic_tfidf_0	0	0.111	0.667	0.889	1	0.246	-0.111	619.778
svm_tfidf_0	0	0.333	0.778	0.889	1	0.278	-0.111	549.889

Table 19: Self Content Politics Metrics

Self_content_social	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
rf_multilingual_0	0.511	0.847	0.977	0.992	0.992	0.544	0.412	207.42
logistic_tfidf_0	0.351	0.763	0.916	0.985	0.985	0.367	0.252	294.817
logistic_multilingual_0	0.489	0.878	0.985	0.992	0.992	0.624	0.389	189.71
svm_multilingual_0	0.504	0.885	0.985	0.985	0.992	0.597	0.405	198.328
svm_bertje_0	0.511	0.84	0.969	0.977	1	0.575	0.412	215.756
nb_tfidf_0	0.191	0.443	0.817	0.916	0.985	0.163	0.092	480.45
svm_tfidf_0	0.321	0.779	0.924	0.985	0.985	0.392	0.221	285.389
rf_bertje_0	0.473	0.824	0.962	0.977	1	0.563	0.374	226.458
logistic_bertje_0	0.573	0.863	0.977	0.985	1	0.571	0.473	191.863
rf_tfidf_0	0.389	0.718	0.908	0.985	1	0.328	0.29	300.366

Table 20: Self Content Social Metrics

Self_content_religion	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_tfidf_0	0.162	0.429	0.749	0.929	0.983	0.161	0.062	513.707
svm_tfidf_0	0.166	0.4	0.737	0.923	0.973	0.146	0.066	531.065
nb_tfidf_0	0.16	0.4	0.711	0.903	0.973	0.108	0.06	549.523
logistic_multilingual_0	0.186	0.45	0.755	0.938	0.989	0.168	0.086	496.295
logistic_bertje_0	0.18	0.431	0.772	0.932	0.985	0.163	0.08	502.222
svm_multilingual_0	0.181	0.429	0.733	0.917	0.979	0.129	0.082	519.408
rf_tfidf_0	0.139	0.373	0.693	0.906	0.982	0.135	0.039	565.746
rf_bertje_0	0.174	0.435	0.764	0.944	0.994	0.195	0.074	496.059
rf_multilingual_0	0.19	0.452	0.76	0.947	0.992	0.195	0.091	495.116
svm_bertje_0	0.162	0.42	0.74	0.909	0.973	0.122	0.062	529.595

Table 21: Self Content Religion Metrics

Self_content_values	Recall_0.1	Recall_0.25	Recall_0.5	Recall_0.75	Recall_0.9	WSS_0.95	ERF_0.1	ATD
logistic_tfidf_0	0.778	0.889	0.889	0.889	1	0.619	0.667	217.222
rf_multilingual_0	0.667	1	1	1	1	0.707	0.556	88
svm_tfidf_0	0.778	0.889	0.889	0.889	1	0.613	0.667	221.667
nb_tfidf_0	0.667	0.778	0.889	0.889	1	0.538	0.556	251.556
logistic_bertje_0	0.778	0.889	1	1	1	0.705	0.667	141.222
svm_multilingual_0	0.778	1	1	1	1	0.734	0.667	79.222
rf_bertje_0	0.556	0.778	1	1	1	0.439	0.444	242.333
svm_bertje_0	0.667	0.889	1	1	1	0.708	0.556	144.556
logistic_multilingual_0	0.889	1	1	1	1	0.734	0.778	69.111
rf_tfidf_0	0.444	0.667	0.889	0.889	0.889	0.419	0.333	329.778

Table 22: Self Content Values Metrics

w1N_SelfEventConnections	**Recall_0.1**	**Recall_0.25**	**Recall_0.5**	**Recall_0.75**	**Recall_0.9**	**WSS_0.95**	**ERF_0.1**	**ATD**
rf_multilingual_0	0.13	0.329	0.608	0.855	0.965	0.075	0.031	638.935
rf_tfidf_0	0.141	0.357	0.691	0.907	0.979	0.135	0.042	570.787
svm_bertje_0	0.134	0.357	0.687	0.9	0.98	0.123	0.035	576.804
rf_bertje_0	0.158	0.385	0.722	0.938	0.987	0.18	0.059	535.8
svm_multilingual_0	0.113	0.306	0.581	0.846	0.961	0.066	0.014	660.57
logistic_multilingual_0	0.124	0.302	0.59	0.862	0.97	0.084	0.025	644.003
nb_tfidf_0	0.148	0.348	0.657	0.867	0.973	0.081	0.049	605.134
logistic_bertje_0	0.154	0.379	0.704	0.934	0.987	0.175	0.055	547.929
logistic_tfidf_0	0.141	0.378	0.693	0.913	0.976	0.143	0.042	560.676
svm_tfidf_0	0.144	0.372	0.68	0.894	0.97	0.09	0.045	576.822

Table 23: w1N_SelfEventConnections Metrics

D Data and Code Statement

D.1 Data Statement

In accordance with ethical guidelines and data privacy regulations, I obtained permission from Jaap Denissen to utilize the dataset for my thesis. As a demonstration of data management practices, all sensitive and identifiable information has been meticulously removed from the dataset, following the appropriate data anonymization procedures. Furthermore, to ensure reproducibility of the simulations, the dataset has been excluded from the version control system through the use of the .gitignore file. The source code and analysis scripts are available in the Git repository, assuming that the necessary data has been acquired independently and with proper authorization.

D.2 Code Statement

For specific guidance and support during the development of the code used in this research, ChatGPT, a language model developed by OpenAI, was consulted. The associated GitHub repository provides comprehensive documentation of the code implementation, including relevant scripts and any modifications made based on the guidance and insights obtained from ChatGPT during the development process. It should be noted that while ChatGPT played a valuable role in providing assistance, the researcher assumes primary responsibility for the overall code development and accuracy.