MASTER THESIS

MuSE-DLF: *Mu*lti-View-*S*emantic *E*nhanced *D*ictionary *L*earning for *F*rame Classification

submitted by

ELIAS CHRISTIAN ANDERLOHR

Submitted to the

Chair for Data Science in the Economic and Social Sciences

within the

Faculty of Business Administration

at University of Mannheim

August 19, 2024

Advisor:

Marlene Lutz, M.Sc. & Ivan Smirnov, Ph.D.

Supervisor:

Prof. Dr. Markus Strohmaier

Declaration of Authorship

I hereby declare that the paper presented is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of other. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet sources. Moreover, I consent to my paper being electronically stored and sent anonymously in order to be checked for plagiarism. I am aware that the paper cannot be evaluated and may be graded failed, if the declaration is not made.

Frankfurt am Main, August 19, 2024

Elias Christian Anderlohr

Abstract *This research advances the field of frame classification in news media through the creation of explainable deep learning models that incorporate semantic role labeling and semantic axis data. We present two novel models: SLMuSE-DLF for single-label frame classification and MuSE-DLF for multi-label frame classification. The SLMuSE-DLF model attained an accuracy of 0.643 and a macro-F1 score of 0.520 on the Media Frames Corpus. Meanwhile, the MuSE-DLF model secured a top-four position in the SemEval-2023 competition with a micro-F1 score of 0.553 and a macro-F1 score of 0.497. Notably, MuSE-DLF is the first model that provides multi-label frame classification along with explainability. It offers both strong classification performance and explainability, as demonstrated through visual techniques for semantic roles and biases. The MuSE-DLF's capability to identify multiple, concurrent frames within a single text more accurately mirrors real-world news framing. This work makes significant contributions to the field of Explainable AI in Natural Language Processing, enhancing our understanding of frame classification in news media while maintaining model explainability.*

Acknowledgements

I would like to express my sincere gratitude to Marlene Lutz, M.Sc. and Ivan Smirnov, Ph.D. for providing important feedback and guidance throughout the development of this thesis. I also thank BwUniCluster 2.0 and the Helix cluster for providing access to high-performance computing resources essential for this work. I acknowledge support by the state of Baden-Württemberg through bwHPC. Furthermore, I acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

Contents

List of Figures

List of Tables

1.

Introduction

Framing is a communication technique often used in fields like politics and media to influence public perception and steer decision-making. It is grounded in the principles of psychology, linguistics, and sociology and has been systematically examined in numerous academic studies [24, 94]. According to Entman [24], framing is the practice of emphasizing certain aspects of reality in communication to shape understanding, problem definition, causal interpretation, moral evaluation and proposed actions. This salience then frames an event and provides a reference point for viewers in which all subsequent information are judged upon [15]. For instance, a political issue may be framed in a way that puts more emphasis on its economic implications than its social consequences, thus influencing the audience's perception of its importance and urgency. This is done through the use of communication frames, which are structures that refer to the way information is presented or communicated to an audience [18, 94]. These frames are constructed through a combination of narratives, symbols, or stereotypes that correspond to a particular ideological framework [49, 46]. Such practices in the media, particularly in sensitive areas such as immigration, can lead to different interpretations by the audience [37]. The effects of framing are illustrated in Figure 1.1, which shows how a single topic can be presented in different ways.

Framing can shape people's perceptions of topics by altering the subjective interpretation of a situation, emphasizing specific arguments while downplaying others, yet leaving the fundamental facts unchanged [19]. To expose the author's subjective influence on people's perception and mitigate the potentially hazardous effects of framing, it is crucial to make the framing visible and measurable [22, 40]. This master's thesis proposes a novel model named the MuSE-DLF model (**Mu**lti-View-**S**emantic **E**nhanced **D**ictionary **L**earning for **F**rame Classification), which enhances frame detection to reveal authors' intentions and identify hidden frames within media narratives [57]. By utilizing semantic information, the model aims to improve the accuracy and explainability of prediction, thus increasing the visibility of these frames and fostering more informed and critical public discourse [57].

Figure 1.1.: *The image showcases the coverage of the same news story by Fox News Latino and Fox News. For the Spanish-speaking audience of Fox News Latino, the headline "In Rare Move, University Grants $22K Scholarship To Undocumented Student" emphasizes the academic achievements and unique nature of the event. This highlights the student's dedication to academics and stresses the significance of being awarded the scholarship. On the other hand, the main Fox News channel uses the headline "Money for Illegals," framing the story with an Economic perspective by focusing on financial aspects and a Legal perspective by highlighting the student's undocumented status. Additionally, Fox News Latino uses more neutral terminology such as "undocumented" to describe the student, while the main Fox News channel chooses the more negative term "illegals" [37] (see Media Matters [65]).*

1.1. Motivation

The motivation for this thesis is rooted in the widespread and often hidden impact of framing within communication. Framing shapes public perception and decision making in multiple fields, such as politics, media, and healthcare [57, 15, 77]. The ability of framing to influence opinions and decisions without people's conscious awareness highlights the need to make these frames visible and measurable [15].

In the field of politics, framing manipulates public sentiment and molds political viewpoints, thus affecting the democratic process by altering perceptions of issues [3]. Media agencies use frame-making tactics to influence the audience's interpretation of events, a phenomenon illustrated by the varied portrayals of immigration by different news outlets [57]. Framing also significantly influences healthcare communication. The presentation of medical information can influence patients' treatment choices, which has important consequences for their health [77]. For example, highlighting survival rates instead of mortality rates can result in different treatment decisions, highlighting the need for visibility [77].

Revealing and making framing visible and measurable offers significant advantages. By uncovering hidden frames in communication, people can become more aware of how their viewpoints and decisions are shaped [51]. This heightened awareness encourages critical thinking, allowing individuals to examine the information they receive more closely and regain control over their decision making processes [17]. Equipping individuals to recognize frames increases their ability to make informed and independent decisions [17]. Ultimately, this leads to a more informed public that is less prone to manipulative tactics and supports clearer, more transparent communication.

1.2. Research Objectives

The primary objective of this thesis is to develop MuSE-DLF (**Mu**lti-View-**S**emantic **E**nhanced **D**ictionary **L**earning for **F**rame Classification), a model developed to contribute to the field of frame detection in Natural Language Processing (NLP), with an emphasis on accuracy and explainability.

We aim to improve the accuracy of frame detection by using enriched semantic information. By integrating semantic information, the model will be capable of detecting frames more accurately in text data. In addition, the model will be designed to handle scenarios with multiple frames, ensuring its flexibility and reliability.

Another key objective is to improve the model's explainability. Our goal is to make the model predictions clear and humanly intelligible to users, thus increasing transparency and confidence in its results. This entails creating techniques to offer explicit insights into the model's decision-making process, promoting a better grasp of the internal mechanisms at play.

In pursuit of these goals, this thesis intends to present a new multi-label model designed to enhance accuracy and explainability in frame classification, addressing open research gaps in frame detection within NLP.

2.

Background

This chapter lays the foundational knowledge needed to understand the research detailed in this thesis. It encompasses discussions on communication framing, explainable artificial intelligence (XAI), and transformer models. Our objective is to establish the theoretical basis essential for understanding this research.

2.1. Introduction to Framing in Communication

Entman [24] describes framing as the process of emphasizing certain aspects of reality in communication in order to influence the perception, problem definition, causal interpretation, moral evaluation, and solution finding of the individual. The term "framing" suggests that the facts presented are embedded in a specific "frame" [54]. Over the years, research has developed independently, resulting in different terms and definitions being attributed to different types of frames.

The term *frame* was initially coined and described in the field of sociology by Bateson [8] as "the spatial and temporal bounding of a set of interactive messages". Goffman [32] later expanded this definition, describing the framework as the organizational factors that control social events and our subjective participation in them [32]. In linguistics, Fillmore [27] coined the term "case frames", which are semantic roles of instruments in a sentence. In psychology, Kahneman and Tversky [49] introduced the term "decision frame", which they defined as "the decision-maker's conception of the acts, outcomes, and contingencies associated with a particular choice".

In 1993, Entman [24] recognised that the definition of "framing" and "frame" was fragmented, and he attempted to restructure the definitions of these terms. Therefore, Entman [24] distinguishes between frames that arise in the mind of the individual and the properties of the messages themselves [94]. Kinder and Sanders [55] in 1996, as well as other authors [21, 86], identified the "double life" of frames, one occurring in communication and the other cognitively [55, 74]. Chong and Druckman [18] summarized the terms as *frame in communication* and *frame in thought*, with

the former also being referred to as *media frame*, which refers to words, images, and phrases, while the latter refers to an intuitive understanding [21, 18].

The most recent and comprehensive definition of framing in research is provided by Sullivan [94], who builds on the work of previous authors to identify three distinct types of frames: semantic, cognitive, and communicative. Semantic frames involve the meaning of words that shape our perception. For instance, the term 'rob' suggests the presence of a thief and a victim, while the term 'sell' indicates a commercial exchange involving a buyer and a seller, thereby shaping our comprehension of the scenario [94, 101]. Cognitive frames refer to how our existing knowledge, such as the concept of private property in the context of robbery, affects our interpretation of events [94]. Communicative frames focus on how the presentation of information can alter its interpretation. For instance, describing an event as a robbery instead of a misunderstanding by emphasizing certain aspects and downplaying others can influence how the event is perceived [94, 24].

The concept of framing, as studied by scholars such as Entman [24], is a multifaceted and interdisciplinary field, with various interpretations in sociology, psychology, and linguistics. This complexity has led to a wide range of views on how data affects our thinking and communication. Sullivan's [94] summary instead provides a clear defined perspective on the framing topic and is thus used for this study. In essence, framing involves the process of emphasizing certain aspects of a situation or event while downplaying others to shape the interpretation and understanding of that situation or event. This process can take place at the semantic level (through the meanings of words), the cognitive level (through our existing knowledge and understanding), and the communicative level (through the way information is presented).

This thesis will adopt the definition of communication frames as described in this chapter. In this document, any mention of *frames* will specifically refer to communication frames.

2.2. Background on Explainable AI

Explainable AI (XAI) is an area of research that seeks to make the decision-making processes of artificial intelligence (AI) systems understandable to humans [5]. The purpose is to improve the interpretability and explainability of machine learning models, which can often be complex and obscure [75]. The significance of XAI is found in its potential to increase trust, transparency, and effectiveness in AI systems in a variety of areas, from the summarization of legal documents [72] to emergency control of the power system [104].

XAI can help users understand why an AI system made a certain decision, which can increase their trust in the system and its results [31]. For example, in a study on legal document summarization, participants completed tasks faster and expressed greater trust in the AI model when explainability features were included [72]. However, it is essential to remember that XAI can also lead to an overdependence on AI advice, sometimes resulting in "blind trust" [85, 90].

Interpretability and explainability are two distinct concepts within the realm of XAI [58, 6]. Interpretability is the degree to which a human can comprehend the cause of a decision made by an AI system [58, 6]. It is related to the inherent understandability of the model itself [2]. Explainability, on the other hand, is often referred to as post-hoc modeling, where only the outcome of a black-box model is justified and explained [28]. It is the extent to which the internal workings of a machine or deep learning system can be explained in human terms [6]. It is about translating the results of a model into a form that humans can comprehend and use to construct a mental model of the AI system [66].

Interpretability is concerned with the transparency of an AI system's decision-making process, while explainability is about making the results of that process understandable to people. Both are essential to help increase trust in AI systems, but focus on different elements of the human-AI relationship.

2.3. Transformer Models

The Transformer model, introduced by Vaswani et al. [98] in the paper "Attention is All You Need", introduced a back then novel architecture that departs significantly from previous sequence learning methods such as recurrent neural networks (RNNs) [50]. Transformer models, including BERT and RoBERTa, focus on tuning their parameters to reduce the cross-entropy loss [64]. This particular loss function is essential during training, especially for activities such as masked language modeling (MLM).

In MLM scenarios, segments of the input text are obscured randomly, with the model's goal being to identify the hidden words using only their surrounding context [64]. This method enhances comprehension of the context and facilitates the acquisition of complex linguistic patterns without requiring specific annotations for each word. MLM frameworks strive to minimize cross-entropy loss by predicting the obscured tokens independently [64], thereby measuring the disparity between two probability distributions [1]. Details of the cross-entropy formula applied in these situations are provided in Figure 2.1.

MLM utilizes the transformer model training by reducing the cross-entropy loss (CE). When the model produces a probability distribution Q and the true distribution is P, the computation of the cross-entropy loss $H(P,Q)$ is as follows:

$$H(P,Q) = -\sum_{x} P(x) \log Q(x) \tag{2.1}$$

Figure 2.1.: *CE $H(P,Q)$ quantifies the average bit count needed to encode events from the actual distribution P utilizing a model that predicts the distribution Q. It aggregates over all potential events x in the dataset, with $P(x)$ representing the actual likelihood of x and $Q(x)$ the probability predicted by the model. This measure reflects how poorly Q approximates P. [44]*

The objective of training is to tune the model's parameters to ensure the predicted distribution Q matches the actual distribution P effectively, thus reducing the CE loss. This optimization process leads the model to make more precise predictions, improving its ability to comprehend language.

3.

Literature Review

This chapter examines the literature relevant for the development of the MuSE-DLF model. By assessing prior research on frame classification and pinpointing gaps in current studies, it lays the groundwork for the contributions of this thesis to frame detection in natural language processing.

3.1. Previous Work on Frame Classification

Investigations into how Framing affects the perception of the audience began in the early 1990s. Primarily in the field of politics, these studies were mainly conducted through manual surveys and investigations. Iyengar's work in the early 1990s was a major breakthrough in the field, as it showed the cognitive effects of how news media present and contextualize an issue or event on audiences and their impact on the perception of political figures [45, 46]. His further analysis of Gulf crisis news coverage revealed how attributes such as news exposure and demographic factors affected opinions on military spending [46]. Nelson, Clawson, and Oxley [71] investigation into public responses to news coverage of a Ku Klux Klan rally provided additional evidence for the effect of framing on public tolerance [71]. Nelson, Clawson, and Oxley [71] discovered that when news stories highlighted free speech, tolerance towards the Ku Klux Klan increased, whereas emphasis on disruption of public order led to decreased tolerance [71].

Early studies in framing research were limited by their reliance on data collected through large-scale surveys. These survey-based approaches could be prone to inaccuracies in participants' responses [20], different interpretations among participants [20], and generally suffer from complexity [25] and high costs [73]. To address this, researchers began to use existing user-generated data, such as articles, comments, and social media interactions. This shift allowed the exploration of political biases and user opinions through new methods, such as language models and network analysis algorithms [76, 68, 105, 81, 7]. This evolution in research methodology reflects a wider move towards utilizing digital data sources in framing studies.

Previous studies have made considerable progress in recognizing political leanings in different texts, yet they often left room for further investigation into the causes of these orientations. This

has led to more detailed research that is attempting to comprehend not only partisanship but also the particular frames used.

Ha's research on Korean news articles concerning the Arab Spring investigated the use of distinct frames, such as *Dictatorship and Oppression* and *Economic Insecurity*, in liberal and conservative media [33]. This approach made the utilization of different frames more evident, illuminating how various media outlets can influence public opinion. Similarly, other scholars have used manually defined frames to recognize political partisanship in US politics [92], and to comprehend users with anti-science or anti-vaccine views on Twitter [82, 87].

Boydstun and colleagues created *The Policy Frames Codebook*, a comprehensive set of frames now known as communicative frames such as *economic* and *morality* frames [12, 94]. This codebook was used to analyze the framing techniques of presidential candidates [93] and to examine changes in news media framing over time, particularly in response to major events such as the September 11 attacks or the 2006 midterm elections [11, 56]. Card et al. [14] development of an annotated dataset covering topics like *death penalty*, *gun control*, and *immigration* enabled researchers to train NLP models for frame prediction at both the sentence and document levels [14, 70, 53].

Boydstun and Gross [12] conducted further research on codebook studies, exploring how agenda setting and framing differ between the two major political parties in the United States. Their results showed that Republicans often frame issues related to taxes from a business point of view, while Democrats usually emphasize the social advantages of taxation [12]. Drawing on the Moral Foundation Theory (MFT) of Haidt and Graham [34], a number of studies have used MFT's predefined moral foundations to classify texts, providing a more detailed understanding of the moral elements embedded in political and social discourse [29, 48, 67, 23, 83, 80].

The approaches discussed in the preceding paragraphs were mainly limited to supporting a single language, mainly English, with a few incorporating languages such as German or Italian. This language-specific focus posed difficulties when analyzing multilingual content. To address this gap, Piskorski et al. [78] created a dataset for the SemEval-2023 Challenge that included a wide range of languages, extending beyond English to French, German, Georgian, Greek, Italian, Polish, Russian, and Spanish. This dataset, annotated with Boydstun et al.'s codebook [12], enabled the exploration of multilingual frame prediction using advanced Transformer-based models. Participants in the challenge employed creative techniques, such as a combination of monolingual and multilingual transformer models [102], a multi-label contrast loss for fine-tuning in multilingual settings [60], and the use of argument data to refine transformer model performance [35].

Other researchers have investigated how to make frame prediction more explainable, using existing linguistic devices or semantic information from words or sentences [57, 53, 103]. These studies align with the concept of "explainability" in Explainable Artificial Intelligence (XAI) by developing post-hoc methods that offer explanations for black-box models. In XAI, explainability refers to the process of justifying and clarifying the decisions made by an AI models, making them more understandable to humans [28, 6]. It is about translating the results of a model into

a form that humans can comprehend and use to construct a mental model of the AI system [66]. Kwak et al. [57] proposed a technique to classify texts based on semantic axes or *semantic axes*, which gives an understanding of the semantic orientation of articles in a single language. The framework recognizes words that have a high semantic weight towards a given axis, which provides a detailed view of how words may affect the perception [57]. Khanehzar et al. [53] used semantic role labeling in sentences to reveal local frame-related information for single-language documents, elucidating the relationship between sentence components and document level frames. Yu [103] concentrated on linguistic devices within German news texts, showing how local textual characteristics can predict the structure of articles in a single language.

3.2. Current Research Gap

The development of Natural Language Processing (NLP) in the field of frame detection has advanced significantly since its early days, which were mainly focused on manual surveys and subjective analysis, as referenced in seminal works [49, 32, 45]. Despite progress and the use of larger datasets to analyze a wider range of media and political biases [76, 7] and the introduction of multilingual and multi-label datasets such as SemEval-2023 [78], there are still open challenges to be addressed. One particular challenge is the development of models that are both explainable and suitable for multi-label classification. Existing models offer explainability but are not capable of multi-label classification. On the other hand, the models that are capable of multi-label classification generally lack features that enable explainability, which is a clear gap in current research.

4.

Research Goals

Having addressed the essential background on communication framing and identified the primary challenges in frame detection as discussed in chapter 3.2, we can now proceed to outline the specific objectives of this study.

1. Enhancement of the FRISS Model with Semantic Data: The primary goal is to develop a novel deep learning model that harnesses semantic role labels along with the semantic orientation of articles to improve frame detection performance. This model will build on top of the FRISS model ([53]), which uses semantic role labels, as well as the semantic inclination of articles using the FrameAxis method of Kwak et al. [57]. These objectives will be pursued in Experiment 1, through a newly proposed model named SLMuSE-DLF (**S**ingle **L**abel **Mu**lti-View-**S**emantic **E**nhanced **D**ictionary **L**earning for **F**rame Classification) (see Chapter 6.2).

2. Development of Multi-Label Frame Classification Capabilities: The second goal is aimed at enhancing the capability of the SLMuSE-DLF model to handle multi-label classification. Such an enhancement is crucial for finer and more accurate frame classification, particularly when texts exhibit overlapping frames. Consequently, the model becomes more versatile in its use and ensures greater accuracy in its classifications. The transition of the SLMuSE-DLF model to a multi-label model called MuSE-DLF (**Mu**lti-View-**S**emantic **E**nhanced **D**ictionary **L**earning for **F**rame Classification) is explored in Experiment 2 (see Chapter 6.3).

3. Ensuring Post-Hoc Explainability: The third objective focuses on ensuring that the MuSE-DLF model remains explainable in line with the definition of XAI outlined in Section 2.2. This entails integrating semantic analysis tools, such as semantic role labeling and semantic inference, into a system that provides post-hoc explainability, making the model's predictions intelligible to users. The model will offer clear and understandable insights into the detection of specific frames, a crucial component in contemporary frame detection research aimed at improving model explainability. This goal is examined in both Experiment 1 (see chapter 6.2) and Experiment 2 (see chapter 6.3).

5.

Methodology

This chapter elaborates on the primary methodologies employed in the thesis: the FRISS model (see Section 5.1), the FrameAxis technique (see Section 5.2), and our proposed MuSE-DLF model (see Section 5.3). Theoretical foundations and practical implementations that are crucial for understanding the adaptations and applications in the following experimental chapters.

5.1. FRISS Model

The FRISS model (**Fr**ame classifier, which is **I**nterpretable and **S**emi-**s**upervised) is designed as a semi-supervised, interpretable multiview framework aimed at predicting media frames, employing semantic role labeling (SRL) to gather and integrate local insights regarding events and their associated actors in news articles [53]. Within this framework, each semantic role embodies a distinct "view." These views are crucial to the framework, offering diverse perspectives or aspects of the information in the text, thereby enhancing the model's comprehension and operational abilities. The architecture of the model includes two main elements: an unsupervised autoencoder to extract latent features and a supervised classification module to determine frames at the document level using these features [53].

5.1.1. Semantic Role Labels (SRL)

Semantic Role Labeling (SRL) is a technique used in NLP that recognizes the functions of words in sentences [89, 69]. Semantic roles have been used in a variety of NLP applications, including information extraction, information retrieval, and other areas, and can improve system performance [100]. Consider the sentence: "Mary gave the book to John." Using SRL, the analysis of this sentence can be seen in Figure 5.1. In this representation, "Mary" is the agent (also called *ARG0*), "gave" is the predicate, and "the book" is the theme (also called *ARG1*). SRL helps to dissect the sentence to understand who is doing what to whom, providing a structured framework for further semantic analysis.

Mary gave the book to John. → Mary$^{\text{ARG0}}$ gave$^{\text{PREDICATE}}$ the book$^{\text{ARG1}}$ to John.

Figure 5.1.: *This figure demonstrates a basic example of semantic role labeling (SRL). In the sentence, different components are color-coded to indicate their semantic roles: the agent (Mary) in blue, the action (gave) in green, and the theme (the book) in red. The predicate can be understood as the event, and the args as the participants.*

5.1.2. Architecture

The architecture of the FRISS model consists out of an unsupervised and supervised component which are both trained in parallel (see Figure 5.2). The FRISS model takes a single article as input. The article is then divided into individual sentences. For each sentence, the SRL spans are extracted using an off-the-shelf SRL model from AllenNLP[1] [30]. Each combination of spans is passed to the unsupervised component.

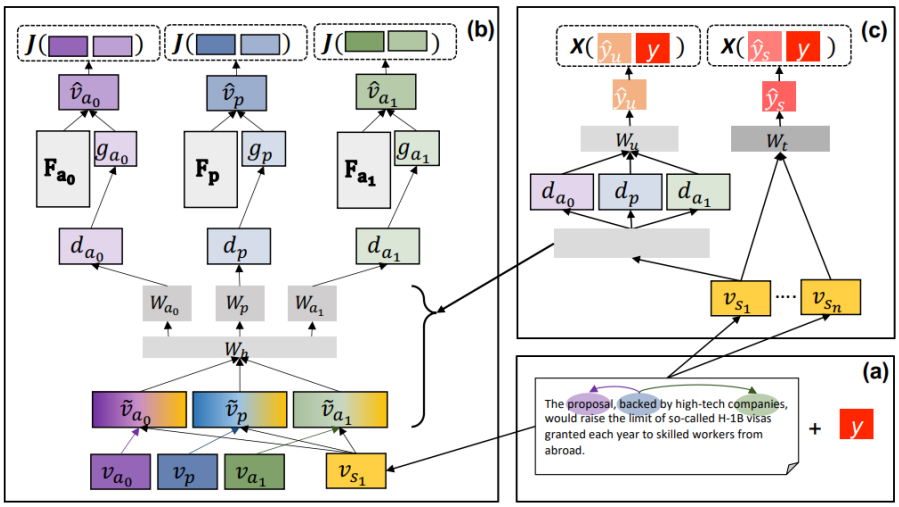

Figure 5.2.: *The FRISS models architecture consisting out of the unsupervised component (b) and supervised component (c) with the input of a document with annotated frame label in (a). (see Khanehzar et al. [53])*

Unsupervised Component

The unsupervised part creates a latent representation for each semantic role, predicting the document-level frame based on the view-specific embedding using a dictionary-learning approach. This section on the unsupervised component refers to part *b)* in Figure 5.2.

Semantic Role Labeling and Embeddings For every *predicate*, *agent*, and *theme*, their span embeddings are derived using a fine-tuned RoBERTa model. Each document undergoes sentence segmentation and is annotated with a transformer-based semantic role labeling model to

[1]The model is trained on the OntoNotes5.0 dataset which consists out of close to 50% news articles.[53]

identify spans for three semantic roles: predicates, agent, and theme. This yields contextualized vector representations for each semantic role span (s_p, s_a, s_t).

Input Representations and Concatenation The procedure is demonstrated using *predicate* as an example, applying similarly to *agent* and *theme*. The embedding for the predicate is represented as v_p:

$$v_p = \frac{1}{|s_p|} \sum_{w \in s_p} v_w, \tag{5.1}$$

where v_w represent the embedding of words in the span s_p. This predicate representation v_p is averaged over all embeddings within the span s_p. Next, we compute the overall sentence embedding v_s:

$$v_s = \frac{1}{|s|} \sum_{w \in s} v_w. \tag{5.2}$$

The average semantic role embedding v_p is concatenated with the sentence representation v_s to form $\tilde{v}_p$ using vector concatenation:

$$\tilde{v}_p = [v_p; v_s], \tag{5.3}$$

where $[;]$ denotes vector concatenation.

Feed-forward Layers and Vector Transformation The concatenated embedding $\tilde{v}_p$ is passed through a series of feed-forward layers. It first goes through a shared feedforward layer W_h with dimensions $2D_w \times D_h$, followed by a role-specific feed-forward layer W_z with dimensions $D_h \times K$. This processing results in a vector l_z:

$$l_z = W_z \text{ReLU}(W_h \tilde{v}_z), \tag{5.4}$$

which is passed in a softmax function, to create a latent representation d_z of the embedding:

$$d_z = \text{Softmax}(l_z). \tag{5.5}$$

Normalization and Reconstruction Subsequently, d_z, a probability vector, is combined with a Gumbel-Softmax sampling approach. The utilization of Gumbel-Softmax facilitates the sampling of discrete categories in a manner that is differentiable and appropriate for gradient-based optimization, essential for the training of neural networks that require discrete decision-making. In the Gumbel-Softmax formula, the temperature parameter τ controls the *"sharpness"* of the distribution, and it is annealed (progressively decreased) during training to transition from a more exploratory behavior (softer, more uniform) to a more exploitative behavior (sharper, more similar to categorical sampling):

$$g_z = \frac{\exp(\log(d_z) + \frac{g}{\tau})}{\sum_f \exp(\log(d_z) + \frac{g}{\tau})}, \tag{5.6}$$

where g is drawn from the Gumbel distribution. The original span embedding $\hat{v}_z$ is then reconstructed by linearly combining the output g_z with a learned, span-specific dictionary F_z:

$$\hat{v}_z = F_z^T g_z. \tag{5.7}$$

Loss Functions The optimization of the reconstructed embedding $\hat{v}_z$ is achieved through specific loss functions designed to improve the correctness of the embeddings. These include contrastive loss and focal triplet loss, each targeting different aspects.

Contrastive Loss The contrastive loss aims to minimize the L2 norm between the original and reconstructed embeddings while maximizing the separation from negatively sampled embeddings. This is formalized as follows:

$$J_z^u(\theta) = \frac{1}{|N_-|} \sum_{v_z^n \in N_-} \max(0, 1 + l_2(\hat{v}_z, v_z) - l_2(\hat{v}_z - v_z^n)). \tag{5.8}$$

For each batch, we generate a set of negative samples, represented as $N\_$, which are utilized for all spans within the current batch.

Focal Triplet Loss The focal triplet loss targets descriptors with lower weights, moving them proportionally further from the most significant descriptor in the reconstruction. Specifically, we select t descriptors with the smallest weights in $\mathbf{g}z$ as additional negative samples, denoted by indices $I = [i_1, i_2, \ldots, i_t]$. The matrix $\mathbf{F}_z^t$ contains these descriptors, with re-normalized weights $\mathbf{g}_z^t = [g_z^{i_1}, g_z^{i_2}, \ldots, g_z^{i_t}]$.

The individual margin for each weight is computed as:

$$m_{it}^z = |M| * (1 - g_{it}^z)^2. \tag{5.9}$$

The focal triplet loss is defined by:

$$J_z^t(\theta) = \frac{1}{|T|} \sum_{i_t \in I} \max(0, m_z^{i_t} + l_2(\hat{v}_z, v_z) - l_2(\hat{v}_z, v_z^{i_t})), \tag{5.10}$$

where T represents the set of descriptors with the smallest weights, I their indices, and m_{it}^z the individual margins.

Combined Loss and Regularization The semantic role-specific loss is determined by adding together the focal triplet loss and the contrastive loss across all spans in $s \in S_z$, and incorporating an orthogonality regularization term:

$$J_z(\theta) = \sum_{s \in S_z} (J_z^u + J_z^t) + \lambda |\mathbf{F}_z \mathbf{F}_z^T - \mathbf{I}|\mathbf{F}^2, \tag{5.11}$$

where λ is a tunable hyper-parameter. The final unsupervised loss is calculated by summing over all spans:

$$J(\theta) = \sum_{z \in p, a_0, a_1} J_z(\theta). \tag{5.12}$$

Supervised Component

The supervised model component aims to improve understanding at the document level by integrating the predictions of the semantic roles of the unsupervised module. This component is divided into two parts, both of which are concurrently developed with the unsupervised model: a Span-based and a Sentence-based Classifier. The supervised component refers to part *c)* in Figure 5.2.

Span-based Classifier The unsupervised module provides predictions at the semantic role level, but the objective here is to predict frame labels at the document level. To achieve this, span-level representations d_s^z are aggregated by averaging both across spans and view:

$$w_u = \frac{1}{Z} \sum_{z \in \{p, a0, a1\}} \frac{1}{|S_z|} \sum_{s \in S_z} d_s^z, \tag{5.13}$$

$$\hat{y}_u = \text{Softmax}(w_u), \tag{5.14}$$

where Z represents the number of views, S_z denotes the set of view-specific spans in the current document, and w_u are the aggregated logits which are then passed through a softmax layer to predict a probability distribution over frames.

Sentence-based Classifier A separate document-level frame prediction is made based on the aggregated sentence-level representations computed in Eq. 5.2. Each sentence embedding is processed through a sequence of feedforward layers and a ReLU nonlinearity to map these to a reduced dimensionality. The results are then averaged across the sentences of the current document and passed through a softmax layer for classification:

$$w_s = \text{ReLU}(W_r v_s), \tag{5.15}$$

$$\hat{y}_s = \text{Softmax}\left(\frac{1}{|S_d|} \sum_{s \in S_d} W_t w_s\right), \tag{5.16}$$

where W_r and W_t are the feed forward layers, and S_d denotes the set of sentences in the document.

Full Loss The supervised and unsupervised components are jointly trained, and the supervised loss $X(\theta)$ is composed of two parts, corresponding to the sentence-based and span-based classifiers:

$$X(\theta) = X(\hat{y}_u, y) + X(\hat{y}_s, y). \tag{5.17}$$

The full model loss combines these supervised components with the unsupervised loss $J(\theta)$, weighted by a hyper-parameter α:

$$L(\theta) = \alpha \times X(\theta) + (1 - \alpha) \times J(\theta), \tag{5.18}$$

where α helps balance the influence of the supervised learning against the unsupervised learning components, facilitating a comprehensive learning framework that leverages both structured and unstructured learning cues from the data.

5.2. FrameAxis

The idea of semantic axes was first proposed by An, Kwak, and Ahn [4] in their publication "SemAxis". These axes represent pairs of contrasting terms, such as "legal – illegal", "clean – dirty", or "fair – unfair", and are used to evaluate the semantic disposition of words and phrases within a text through word embeddings. FrameAxis, a text analysis approach evolved from the "SemAxis" technique by Kwak et al. [57], coins the term "microframe" for semantic axis and introduces two additional metrics[2]: microframe bias (termed as semantic axis bias hereafter) and microframe intensity (referred to as semantic axis intensity). Figure 5.3 illustrates how these metrics capture different aspects of semantic framing. Semantic axis bias evaluates the text's inclination towards one end of the semantic axis, uncovering the main semantic trend. On the other hand, semantic axis intensity measures the frequency of a semantic axis in the text, underscoring the importance and density of particular semantic topics.

In the context of discussing immigration issues within an 'illegal – legal' semantic axis, the framing bias would measure the extent to which the text emphasizes 'illegal' as opposed to 'legal' elements of immigration. At the same time, the framing intensity would evaluate how strongly the text focuses on these legal or illegal viewpoints compared to other possible frames.

Utilizing word embeddings, FrameAxis identifies the similarities between words and semantic axes. This technique offers an framework for examining the semantic trends that support textual narratives. This study will employ these tools to detect the subtle semantic distinctions of media frames, thereby improving the model's predictive accuracy and explainability.

[2] In this study, we will exclusively use the term *semantic axis* instead of *microframe*, as the latter may cause confusion given our objective to predict *frames*

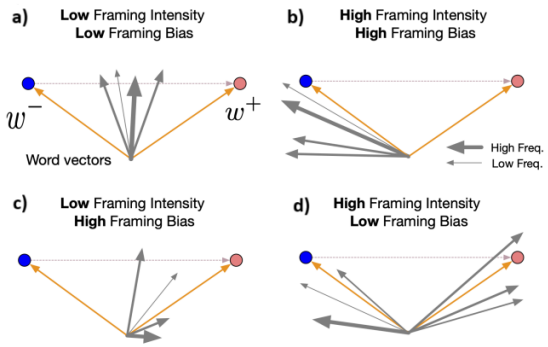

Figure 5.3.: *Depiction of semantic axis intensity and bias adapted from Kwak et al. [57]. (a) Low intensity, low bias: neutral approach with rare usage of axis-related terms. (b) High intensity, high bias: pronounced focus on one end of the spectrum, signifying a clear framing direction. (c) Low intensity, high bias: subtle inclination towards one end without strong emphasis. (d) High intensity, low bias: frequent usage of terms tied closely to both ends, indicating a balanced yet emphatic coverage. Blue and orange circles denote pole word vectors (w- and w+), gray arrows represent corpus word vectors, with arrow thickness showing frequency.*

5.2.1. Measurements

The two FrameAxis measurements, semantic axis bias and intensity, are computed as follows. Initially, a semantic axis must be established. This consists of constructing a *positive* ($\mathbf{v}_w^+$) and a *negative* ($\mathbf{v}_w^-$) embedding. The final semantic axis $\mathbf{v}_f$ is then formed by subtracting the negative embedding from the positive one, expressed as $\mathbf{v}_f = \mathbf{v}_w^+ - \mathbf{v}_w^-$.

Word Contribution: To assess the bias and intensity of the semantic axis, it is important to evaluate the semantic orientation of each individual word in each document in the corpus relative to the semantic axis. This calculation involves determining the word contribution $\mathbf{c}_f^w$, defined as the cosine similarity between a word's embedding $\mathbf{v}_w$ and the semantic axis $\mathbf{v}_f$:

$$\mathbf{c}_f^w = \frac{\mathbf{v}_w \cdot \mathbf{v}_f}{\|\mathbf{v}_w\| \|\mathbf{v}_f\|}. \tag{5.19}$$

Microframe Bias (Semantic Axis Bias): Microframe bias, referred to in this study as semantic axis bias, expands upon the idea of word contribution, which assesses the semantic orientation of specific words. In contrast, semantic axis bias aggregates this information for each document in the corpus to determine whether the whole document leans positively or negatively towards the semantic axis. A positive bias, $\mathbf{B}_f^t$, nearing 1, indicates that the corpus predominantly uses words that favorably align with the set semantic axis. Conversely, a negative bias nearing -1 indicates the opposite effect. The calculation of $\mathbf{B}_f^t$ is as follows:

$$\mathbf{B}_f^t = \frac{\sum_{w \in t} (n_w \mathbf{c}_f^w)}{\sum_{w \in t} n_w}, \tag{5.20}$$

where n_w represents the frequency of the word w in the corpus t.

Microframe Intensity (Semantic Axis Intensity): To delve deeper into the role of the semantic axis within the corpus, we utilize the measurement microframe intensity, which we denote in this work as semantic axis intensity. This metric assesses the prominence of a semantic axis in the corpus, with higher values signifying more significance. The calculation of intensity involves the second moment of word contributions in relation to the corpus-wide average bias:

$$\mathbf{I}_f^t = \frac{\sum_{w \in t}(n_w(\mathbf{c}_f^w - \mathbf{B}_f^t)^2)}{\sum_{w \in t} n_w}, \tag{5.21}$$

where n_w represents the frequency of the word w in the corpus t and $\mathbf{B}_f^t$ denotes the overall bias of corpus t.

Semantic Axis word-level shift: To investigate the influence of individual words on the semantic axis f with respect to $\mathbf{S}_w^t(B_f)$ and semantic axis intensity $\mathbf{S}_w^t(I_f)$, a method known as word-level shift was employed. This method involves assessing the difference in bias or intensity between a target corpus and a background corpus through the following equations:

$$\mathbf{S}_w^t(B_f) = \frac{n_w c_f^w}{\sum_{w \in t} n_w}, \tag{5.22}$$

and

$$\mathbf{S}_w^t(I_f) = \frac{n_w(c_f^w - \mathbf{B}_f^T)^2}{\sum_{w \in t} n_w}, \tag{5.23}$$

where w represents the specific word, c_f^w indicates the contribution of word w to semantic axis f, n_w denotes the frequency of word w in corpus t, and $\mathbf{B}_f^T$ is the bias of the background corpus. The background corpus is created by selecting contrasting documents. For example, if the articles are positive about immigration, the background corpus would include articles that are not positive about immigration.

5.2.2. Semantic Axis Design

Our methodology for semantic axis design is underpinned by the Extended Moral Foundation Theory (eMFT), as elaborated by Hopp et al. [38]. This theory extends the seminal principles initially posited by Haidt and Graham [34]. It methodically outlines moral reasoning across five separate foundations: *care/harm*, *fairness/cheating*, *loyalty/betrayal*, *authority/subversion*, and *sanctity/degradation* (see Table 5.1 for descriptions of these moral foundations) Haidt and Graham [34]. Each of these foundations comprehensively embodies the dualistic nature of morality, which encompasses both virtues and vices Haidt and Graham [34].

To convert these foundations into semantic axis for textual analysis, we used a methodology that goes beyond the traditional use of single-word antonyms to describe semantic dimensions. In line with recent advances in semantic analysis [16, 67], we curated a list of words for each

Moral Foundation	Explanation
Care/Harm	Encompassing feelings of sympathy, compassion, and care
Fairness/Cheating	Encompassing ideas of rights and justice
Loyalty/Betrayal	Upholding moral duties of patriotism and 'us versus them' mentality
Authority/Subversion	Pertaining to issues of tradition and the preservation of social order
Sanctity/Degradation	Relating to moral repulsion and spiritual issues concerning the body

Table 5.1.: *The five Moral Foundations of the MFT by Haidt and Graham [34]*

moral foundation, thus augmenting semantic granularity. Each foundation was divided into two distinct categories: *virtue* and *vice*. For each category, we identified the 40 words (referred to as key words) with the highest likelihood of representing the respective moral sentiment, based on the dataset from [38]. Table 5.2 shows the top five key words for each moral sentiment (*vice* and *virtue*) and moral foundation combination[3].

Moral Foundation	Category	Top 5 Words
Authority/Subversion	Vice	protested, rebellion, protesting, revenge, rage
Authority/Subversion	Virtue	recommended, authority, revive, promising, embrace
Care/Harm	Vice	tortured, cruel, harsh, hostility, killed
Care/Harm	Virtue	compassion, welcomed, friendly, treat, friendship
Fairness/Cheating	Vice	rigged, undermining, punished, steal, inability
Fairness/Cheating	Virtue	freedoms, fair, honored, wealthy, integrity
Loyalty/Betrayal	Vice	betrayal, fearful, stealing, hateful, rejection
Loyalty/Betrayal	Virtue	loyalty, loves, improvement, solidarity, willingness
Sanctity/Degradation	Vice	raping, exploit, rape, victimized, distrust
Sanctity/Degradation	Virtue	praise, celebrating, virtues, strengthening, respected

Table 5.2.: *Top 5 Terms in Each Category for the Five Moral Foundations*

5.2.3. Calculation of FrameAxis values using Contextualized Embedding

The creators of FrameAxis utilized static GloVe embeddings to determine semantic axis bias and intensity. Other researchers propose that employing pre-trained transformer models such as BERT might eliminate unintended biases [10, 95] and generally enhance the reliability of FrameAxis outcomes [16]. Consequently, we employ a fine-tuned RoBERTa model for computing the FrameAxis metrics.

Semantic Axis Calculation Let $K = k_1, k_2, \ldots, k_n$ denote a set of keywords, where each k_i is lemmatized to its base form b_i. The methodology begins with the identification of the sentences. For each $k_i \in K$, we define I^{k_i} as the set of all sentences containing any morphological variant of

[3]The complete list of all 40 words is available in the Appendix under C.1

b_i. This step ensures that we capture all relevant contexts in which the keyword or its variants appear.

Following sentence identification, we proceed with tokenization and embedding. For each sentence $s \in I^{k_i}$, we tokenize s into a sequence of tokens $\mathbf{T}_s = [t_1, t_2, \ldots, t_m]$. We then generate contextual embeddings $\mathbf{E}_s = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m]$ using a fine-tuned RoBERTa model, where $\mathbf{e}_j \in \mathbb{R}^d$ and d is the dimensionality of the embedding space. This process converts the text data into a comprehensive, context-sensitive numerical format represented as transformer embeddings.

The next phase involves the extraction of keywords. For each sentence s, we identify the set of indices $J_s = j : t_j$ corresponds to b_i in s. We then extract the relevant embeddings $\mathbf{e}_j : j \in J_s$. To account for cases where a keyword may be tokenized into multiple tokens, we compute the average embedding for k_i in sentence s, in the following way:

$$\mathbf{v}_s^{k_i} = \frac{1}{|J_s|} \sum_{j \in J_s} \mathbf{e}_j. \tag{5.24}$$

This step ensures that we have a single, representative embedding for each occurrence of the keyword in a sentence, regardless of how it was tokenized.

The final step in our methodology is the computation of the overall keyword embedding. We calculate the embedding $\mathbf{v}^{k_i}$ for keyword k_i by averaging across all sentences in I^{k_i}. The formula for this calculation is:

$$\mathbf{v}^{k_i} = \frac{1}{|I^{k_i}|} \sum_{s \in I^{k_i}} \mathbf{v}_s^{k_i}, \tag{5.25}$$

where $|I^{k_i}|$ denotes the cardinality of the set I^{k_i}. This final averaging step produces a single, contextual embedding that represents the keyword's usage across the entire corpus.

Once we have generated a contextual embedding for each keyword $k \in K$, we can proceed with the creation of the semantic axis. Initially, we divide the list of keywords K into two groups. Let's define two subsets for our keyword set K: $P \subset K$, which signifies virtue-related keywords, and $N \subset K$, which signifies vice-related keywords. With the keyword embeddings $\mathbf{v}_{k_i}$ obtained through our technique, we can now determine aggregate embeddings for virtue and vice as outlined:

$$\mathbf{v}_{\text{virtue}} = \frac{1}{|P|} \sum_{k_i \in P} \mathbf{v}_{k_i}, \tag{5.26}$$

$$\mathbf{v}_{\text{vice}} = \frac{1}{|N|} \sum_{k_i \in N} \mathbf{v}_{k_i}, \tag{5.27}$$

where, $|P|$ and $|N|$ denote the cardinalities of the virtue and vice keyword sets, respectively. From these aggregate embeddings, we can derive a semantic axis vector $\mathbf{v}_f$, which represents the semantic difference between virtue and vice in our embedding space:

$$\mathbf{v}_f = \mathbf{v}_{\text{virtue}} - \mathbf{v}_{\text{vice}} \tag{5.28}$$

This methodology provides a robust mechanism to encode the context of the keywords into the semantic axis creation.

Word Contribution The method for determining word contributions starts by iterating through each sentence s_n indexed by N within the dataset. For each sentence s_n in our corpus, we generate token embeddings $\mathbf{E}s = [\mathbf{e}s1, \mathbf{e}s2, \ldots, \mathbf{e}sm]$, where m is the number of tokens in the sentence. To maintain the integrity of our analysis, we identify and exclude embeddings of stop words from $\mathbf{E}_s$. This exclusion is performed after the tokenization and embedding process to preserve the original contextual information.

Following the removal of stop word embeddings, the remaining embeddings in $\mathbf{e}_{sj}$ in sentence s_n are analyzed to determine their contribution to each semantic axis $\mathbf{v}_f$. For every word embedding $\mathbf{e}_{sj}$ still present in d_n, the cosine similarity with the semantic axis embedding $\mathbf{v}_f$ is calculated. This process is repeated for each word embedding in every sentence

$$\text{cosine\_similarity}(\mathbf{e}_{sj}, \mathbf{v}_f) = \frac{\mathbf{e}_{sj} \cdot \mathbf{v}_f}{\|\mathbf{e}_{sj}\| \|\mathbf{v}_f\|}. \tag{5.29}$$

This approach assesses the impact of each non-stop word on the semantic direction established by the semantic axis.

Semantic Axis Bias and Intensity Measurement Building on the methods described in an earlier chapter, we compute the bias and intensity for each semantic axis according to the measurements specified in Equations 5.20 and 5.21. These calculations are performed using the contextualized embeddings and the word contributions contextualized from the preceding analysis. This method guarantees precise evaluation of both bias and intensity, effectively reflecting the semantic direction and the weight of words in relation to each semantic axis in the dataset.

5.3. MuSE-DLF

This chapter explores the design and evolution of the MuSE-DLF model, a model for frame classification in NLP. MuSE-DLF stands for *Mu*lti-View-*S*emantic *E*nhanced *D*ictionary *L*earning for *F*rame Classification, and its single-label predecessor is called SLMuSE-DLF, which stands for *S*ingle *L*abel *Mu*lti-View-*S*emantic *E*nhanced *D*ictionary *L*earning for *F*rame Classification. The chapter is organized to first discuss the vital preprocessing steps needed for the model and dataset, and then introduce SLMuSE-DLF, the single-label predecessor to MuSE-DLF. This sequence emphasizes the model's evolution from the FRISS framework and its incorporation of the FrameAxis component, leading to the development of the multi-label capable MuSE-DLF model.

5.3.1. Preprocessing

To prepare data for the SLMuSE-DLF and MuSE-DLF model, it is crucial to carry out several preprocessing steps to ensure the data is correctly formatted and contains the necessary information for both unsupervised and supervised elements. These steps include the one-hot encoding for the SLMuSE-DLF and multi-hot encoding for the MuSE-DLF of document-level frames, sentence segmentation, semantic role extraction, and the computation of FrameAxis values.

Initially, document-level frames undergo either one-hot or multi-hot encoding. The SLMuSE-DLF model necessitates one-hot encoded document-level frames, whereas the MuSE-DLF model requires multi-hot encoded ones. These methods generate a binary vector for each document, with each vector element corresponding to a particular frame. A value of 1 indicates a frame's presence, while a value of 0 indicates its absence. This encoding technique allows the model to effectively encode the document-level frame and make it applicable for the models.

After encoding, the text of each document is broken down into individual sentences. Each segmented sentence is subsequently tagged with the document-level frames from its source document, preserving the contextual association between the sentences and their overall frames. This forms a dataframe where each document consists of a sublist containing all its sentences.

Afterward, all semantic roles within each sentence are identified. Semantic role labeling determines the main verb (predicate) and its related arguments, such as agent (also called ARG0) and theme (also called ARG1). Since a sentence can have multiple combinations of semantic roles and a word can fulfill multiple roles, the preprocessing step ensures that all combinations of semantic roles are identified. Each combination must contain a non-empty predicate and at least one non-empty argument (either agent or theme), thus guaranteeing that each semantic role pair has enough information for the model to process effectively.

Furthermore, the FrameAxis metrics, which include semantic axis bias and intensity, are calculated for each sentence following the procedure described in Section 5.2.3.

Through these preprocessing steps, the data is transformed into a format necessary for the SLMuSE-DLF and MuSE-DLF models, ensuring that it is enriched with the required features for frame classification.

5.3.2. Model Architecture

The SLMuSE-DLF model precedes the MuSE-DLF model, with both models being adaptations of the FRISS model and incorporating FrameAxis to improve frame classification. They share similar data flow structures and integrate both unsupervised and supervised components to analyze text thoroughly for frame classification. These models build on the FRISS model [53] by introducing new elements and modifying existing ones to improve performance, as shown in Figure 5.4.

A comprehensive description of the SLMuSE-DLF model is provided in section 5.3.3, whereas section 5.3.4 details the modifications of the MuSE-DLF model compared to the SLMuSE-DLF model.

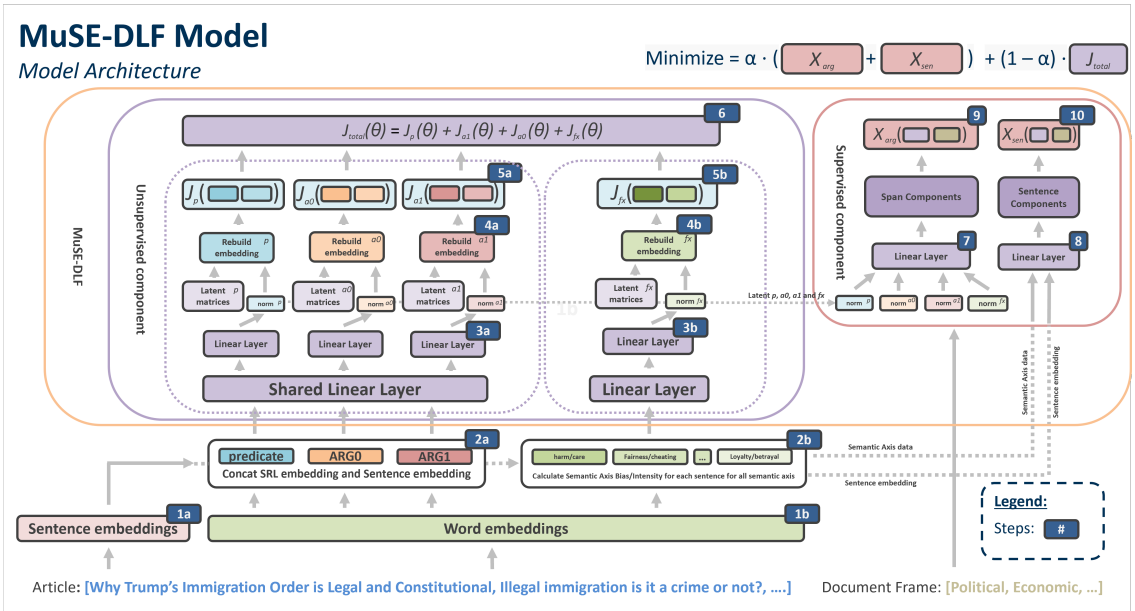

Figure 5.4.: *The MuSE-DLF and SLMuSE-DLF model architecture incorporates both unsupervised and supervised components for comprehensive text processing. Initially, sentence (Step 1a) and word embeddings (Step 1b) are generated using a transformer model. The word embeddings for the predicate, ARG0 (agent), and ARG1 (theme) are extracted and concatenated with the sentence embedding (Step 2a). In parallel, the FrameAxis bias and intensity values for a set of semantic axis are calculated and concatenated with the sentence embedding, resulting in additional dimensions (Step 2b). These combined embeddings are then processed through shared and individual linear layers to reduce dimensions and create latent representations (Steps 3a and 3b). Embeddings are reconstructed back to their original dimensions (Steps 4a and 4b) and compared to the original embeddings (Steps 5a and 5b), contributing to the total unsupervised loss J_{total} (Step 6). For the supervised component, latent representations are averaged over all sentences (Step 7) and used to predict document-level frames (Step 8), which are then compared to the true classes (Steps 9 and 10). Training aims to balance unsupervised and supervised losses, ensuring effective embedding reconstruction and accurate frame prediction.*

5.3.3. SLMuSE-DLF: Single-Label MuSE-DLF Model

The model handles the input by splitting it into groups of documents, where each document contains several sentences (refer to Step 1a and Step 1b in 5.4). Each sentence within these documents is associated with a set of semantic role arguments (predicate, agent, and theme) and its corresponding Semantic Axis values. Let a_i be the i-th article in the current batch with a maximum m sentences each. Initially, the model creates sentence and word embeddings using a transformer model like RoBERTa. These embeddings act as the basic representations for later stages of processing.

For every sentence s, the embeddings are retrieved for the predicate, agent, and theme. The process of obtaining these embeddings is similar to that used to extract the contextualized embedding for the Semantic Axis measurement (see Section 5.2.3). This involves initially tokenizing each

sentence s within the article a into a sequence of tokens $\mathbf{T}_s = [t_1, t_2, \ldots, t_n] \in \mathbb{R}^n$, where n is the number of tokens in the sentence. These tokens are then converted into contextual embeddings $\mathbf{E}_s = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n] \in \mathbb{R}^{n \times d}$, where d is the embedding dimension.

Semantic Role and Semantic Axis Embedding Extraction

Leveraging the derived word and sentence embeddings, we proceed to compute the semantic role and semantic axis information (see Step 2a and Step 2b in 5.4). For each occurrence of a semantic role $z \in \{predicate, agent, theme\}$ in the sentences, we locate the indices in $\mathbf{T}s$ that correspond to the tokens representing z, denoted as I^z. We aggregate the embeddings at these indices, averaging them to form a single contextual vector $\mathbf{v}_s^z \in \mathbb{R}^d$ for each occurrence of the role z:

$$\mathbf{v}_s^z = \frac{1}{|I^z|} \sum_{i \in I^z} \mathbf{e}_i, \tag{5.30}$$

where I^z is the set of indices in $\mathbf{T}_s$ corresponding to the tokens representing z, $\mathbf{e}_i \in \mathbb{R}^d$ is the embedding of the token at index i, and $|I^z|$ is the size of the set I^z.

The sentence embedding, represented by $\mathbf{v}_s \in \mathbb{R}^d$, is derived by averaging the embeddings of non-padded tokens within a sentence[4]:

$$\mathbf{v}_s = \frac{1}{\sum_{i=1}^{|\mathbf{T}s|} \mathbf{1}[t_i \neq \text{PAD}]} \sum_{i=1}^{|\mathbf{T}_s|} \mathbf{e}_i \cdot \mathbf{1}[t_i \neq \text{PAD}], \tag{5.31}$$

where $|\mathbf{T}_s|$ is the total number of tokens in the sentence, $\mathbf{e}i \in \mathbb{R}^d$ is the embedding for the i-th token, and $\mathbf{1}[t_i \neq \text{PAD}]$ is an indicator function that equals 1 when the i-th token is not a padding token and 0 otherwise. The denominator $\sum_{i=1}^{|\mathbf{T}s|} \mathbf{1}[t_i \neq \text{PAD}]$ calculates the number of non-padded tokens, ensuring that the average is computed only over meaningful tokens.

These extracted embeddings are then concatenated with the embedding of the sentence $\mathbf{v}_s$, forming a comprehensive representation that includes both local and contextual information (Step 2a). This step, derived from the FRISS model, remains unchanged and is performed for $\mathbf{v}_s^p$, $\mathbf{v}_s^a$, and $\mathbf{v}_s^t$:

$$\tilde{\mathbf{v}}_s^p = [\mathbf{v}_s^p; \mathbf{v}_s], \quad \tilde{\mathbf{v}}_s^a = [\mathbf{v}_s^a; \mathbf{v}_s], \quad \tilde{\mathbf{v}}_s^t = [\mathbf{v}_s^t; \mathbf{v}_s] \tag{5.32}$$

where $\mathbf{v}_s^p, \mathbf{v}_s^a, \mathbf{v}_s^t, \mathbf{v}_s \in \mathbb{R}^d$ are the embeddings of the predicate, agent, theme, and the original sentence, respectively. $[;]$ represents the concatenation of vectors, resulting in $\tilde{\mathbf{v}}_s^p, \tilde{\mathbf{v}}_s^a, \tilde{\mathbf{v}}_s^t \in \mathbb{R}^{2d}$.

Simultaneously with Step 2a, in Step 2b the model calculates the Semantic Axis bias (B_s^f) and intensity (I_s^f) metrics for all semantic axes f and each sentence $s \in a_i$, as described in Section 5.2.3. In the SLMuSE-DLF model, the Semantic Axis method involves the usage of five semantic axes, each contributing two scalar values, the semantic axis bias and intensity, resulting in a total of ten new values that are concatenated to the sentence embedding, following the logic applied

[4]It was found that averaging the embeddings was more effective during model development than using the *CLS* embedding from the sentence.

to the predicate, agent, and theme embeddings above (see Equation 5.32). For each semantic axis f, we calculate two values: the semantic axis bias (B_s^f) and the semantic axis intensity (I_s^f). Consequently, for five semantic axes, the resulting values form a vector $\mathbf{v}_s^{fx} \in \mathbb{R}^{10}$:

$$\mathbf{v}_s^{fx} = \left\{ \begin{array}{ll} \mathbf{B}_s^{\text{care/harm}}, & \mathbf{I}_s^{\text{care/harm}}, \\ \mathbf{B}_s^{\text{fairness/cheating}}, & \mathbf{I}_s^{\text{fairness/cheating}}, \\ \mathbf{B}_s^{\text{loyalty/betrayal}}, & \mathbf{I}_s^{\text{loyalty/betrayal}}, \\ \mathbf{B}_s^{\text{authority/subversion}}, & \mathbf{I}_s^{\text{authority/subversion}}, \\ \mathbf{B}_s^{\text{sanctity/degradation}}, & \mathbf{I}_s^{\text{sanctity/degradation}} \end{array} \right\}. \tag{5.33}$$

These ten values are then concatenated with the sentence embedding $\mathbf{v}_s$ as follows:

$$\tilde{\mathbf{v}}_s^{fx} = [\mathbf{v}_s^{fx}; \mathbf{v}_s] \in \mathbb{R}^{d+10}, \tag{5.34}$$

where $\mathbf{v}_s^{fx} \in \mathbb{R}^{10}$ denotes the vector of semantic axis bias and intensity values, $[;]$ signifies vector concatenation, and $\mathbf{v}_s \in \mathbb{R}^d$ stands for the sentence embedding. It should be noted that the SLMuSE-DLF model can accommodate various types of Semantic Axis inputs, which means the model can work with the standard 5 semantic axes, additional ones, or entirely different sets.

Unsupervised Component: Dimensionality Reduction and Latent Representation

The combined embeddings from Steps 2a and 2b are subsequently processed through shared and individual linear layers to reduce their dimensionality and create latent and sparse representations (Steps 3a and 3b). Initially, the embeddings are processed through a common feed-forward layer utilizing GELU activation[5], and then through view-specific feed-forward layers. This reduction in dimensionality, crucial for uncovering significant latent features while minimizing computational costs, is derived from the FRISS model with some minor modifications.

Let h denote the hidden size dimension, and C the number of classes to predict. Due to the different sizes in the dimensions of $\tilde{\mathbf{v}}_s^{fx} \in \mathbb{R}^{d+10}$ compared to the three semantic role embeddings $\tilde{\mathbf{v}}_s^p, \tilde{\mathbf{v}}_s^a, \tilde{\mathbf{v}}_s^t \in \mathbb{R}^{2d}$, the processing is split apart. Thus, $\tilde{\mathbf{v}}_s^p$, $\tilde{\mathbf{v}}_s^a$, and $\tilde{\mathbf{v}}_s^t$ are each processed through a shared layer $\mathbf{W}_h \in \mathbb{R}^{2d \times h}$ before passing through three individual linear layers $\mathbf{W}_{h1}, \mathbf{W}_{h2}, \mathbf{W}_{h3} \in \mathbb{R}^{h \times h}$, and lastly into another individual output layer $\mathbf{W}_{zo} \in \mathbb{R}^{h \times C}$, resulting in $\mathbf{l}^z \in \mathbb{R}^C$.

It is noteworthy that subsequent to each linear layer in both the $ilde\mathbf{v}_s^{fx}$ and $ilde\mathbf{v}_s^z$ pathways, the data is subjected to a series of operations: initially, it goes through a dropout layer, followed by a GELU activation function, and ultimately a batch normalization layer, before proceeding to the next linear layer.

In contrast, $\tilde{\mathbf{v}}_s^{fx}$ undergoes processing through multiple distinct linear layers, bypassing the shared linear layer $\mathbf{W}_h$. The initial processing is given by:

[5]The initial implementation of FRISS used ReLU for activation, however, in our configuration, GELU yielded superior results. Additionally, other studies have highlighted that GELU is particularly effective with embeddings generated by Transformers [91].

$$\hat{\mathbf{l}}_s^{fx} = \text{GELU}(\mathbf{W}_{fx1}\tilde{\mathbf{v}}_s^{fx}), \tag{5.35}$$

where $\mathbf{W}_{fx1} \in \mathbb{R}^{(d+10)\times h}$ is the initial linear layer used to reduce dimensionality. Subsequently, $\hat{\mathbf{l}}_s^{fx}$ is processed through three individual hidden linear layers $\mathbf{W}_{fxh1}, \mathbf{W}_{fxh2}, \mathbf{W}_{fxh3} \in \mathbb{R}^{h\times h}$, and a final output layer $\mathbf{W}_{fxo} \in \mathbb{R}^{h\times C}$ to produce $\mathbf{l}_s^{fx}$. In order to maintain conciseness, we do not include the detailed mathematical expressions for this part of the process. During this entire procedure, we apply the GELU activation function [5].

Afterward, $\mathbf{l}_s^{fx}$ and $\mathbf{l}^z$ are processed using a softmax function to produce $\mathbf{d}_s^{fx}$ and $\mathbf{d}_s^z$ respectively, both in $\mathbb{R}^C$:

$$\mathbf{d}_s^z = \text{softmax}(\mathbf{l}^z), \quad \mathbf{d}_s^{fx} = \text{softmax}(\mathbf{l}_s^{fx}). \tag{5.36}$$

This process creates a compact, informative encoding of the input by progressively reducing the dimensionality from $2d$ to h, and finally to C. Additionally, we generate $\mathbf{g}_s^z$ and $\mathbf{g}_s^{fx}$ by drawing samples from the Gumbel-Softmax distribution (refer to Section 5.1.2).

Unsupervised Component: Loss Computation

After dimensionality reduction, the latent representations are restored to their original dimensions. In Steps 4a and 4b, this restoration process involves the generation of span-specific dictionary representations. The semantic role embeddings $\tilde{\mathbf{v}}_s^z \in \mathbb{R}^{2d}$ are reconstructed and compared with the initial embeddings $\mathbf{v}_s^z$ (Step 4a), while simultaneously the Semantic Axis embeddings $\tilde{\mathbf{v}}_s^{fx} \in \mathbb{R}^{d+10}$ are similarly reconstructed and compared against their initial Semantic Axis values (Step 4b).

The reconstruction process can be expressed as:

$$\hat{\mathbf{v}}_s^z = (\mathbf{F}^z)^{\mathbf{T}}\mathbf{d}_s^z, \quad \hat{\mathbf{v}}_s^{fx} = (\mathbf{F}^{fx})^{\mathbf{T}}\mathbf{d}_s^{fx}, \tag{5.37}$$

where $\mathbf{F}^z \in \mathbb{R}^{d\times C}$ and $\mathbf{F}^{fx} \in \mathbb{R}^{10\times C}$ are learnable dictionary matrices, and $\hat{\mathbf{v}}_s^z$ and $\hat{\mathbf{v}}_s^{fx}$ are the reconstructed embeddings.

The unsupervised loss is derived by computing the reconstruction loss, which comprises the L2 norm between the original and reconstructed embeddings, while also enhancing the distinction from negatively sampled embeddings (Steps 5a and 5b). This approach minimizes the reconstruction error while increasing the distance to the negatively sampled embeddings, ensuring distinctiveness. The unsupervised loss computation for the embeddings (Step 5a) and for the Semantic Axis values (Step 5b) follows the established FRISS model methodology (see Section 5.1.2 for a detailed explanation of the loss calculation). The total loss $J(\theta)$ is obtained by first calculating the reconstruction errors for each sentence and its spans for all semantic roles and Semantic Axis values. After computing these losses for all sentences and all four views (predicate, agent, theme, and Semantic Axis), they are summed to form a single overall loss, which is then scaled by the number

of entries, thus normalizing the unsupervised loss compared to the initial FRISS implementation[6] (refer to Step 6 in Figure 5.4):

$$J(\theta) = \frac{1}{|Z| \cdot m} \sum_{z \in \{p,a,t,fx\}} \left(\sum_{s \in a_i} J_s^z(\theta) \right), \tag{5.38}$$

where $J_s^z(\theta)$ computes the reconstruction errors for view z and sentence s, with $|Z|$ representing the number of views (four in this case: predicate, agent, theme, and Semantic Axis), a_i is the current article i, and m represents the number of sentences per article.

Supervised Component

In the supervised component, the model consists of a span-based classifier (Step 7 and Step 9) and a sentence-based classifier (Step 8 and Step 10); together, they predict the document-level frame.

Span-based Classifier: The span-based classifier relies on the latent representations $\mathbf{d}_s^z \in \mathbb{R}^C$ created by the unsupervised components. These representations are first averaged over all sentences within each article (Step 6). At this point, the latent Semantic Axis measurements are also integrated along with the latent predicate, agent, theme representations, altering the original FRISS method:

$$\mathbf{y}_{span} = \frac{1}{|Z|} \sum_{z \in \{p,a,t,fx\}} \left(\frac{1}{|S_i^z|} \sum_{s \in S_i^z} \mathbf{d}_s^z \right) \in \mathbb{R}^C, \tag{5.39}$$

where $|Z| = 4$ indicates the total number of views, and S_i^z denotes the collection of spans specific to view z for the current article i.

Sentence-based classifier: The sentence-based classifier predicts the document-level frame using the embeddings of all sentences within the document. For each sentence s, we have a sentence embedding $\mathbf{v}_s \in \mathbb{R}^d$ and a Semantic Axis vector $\mathbf{v}_s^{fx} \in \mathbb{R}^{10}$. These are concatenated to form $\tilde{\mathbf{v}}_s = [\mathbf{v}_s; \mathbf{v}_s^{fx}] \in \mathbb{R}^{d+10}$, where $[;]$ denotes concatenation. Each article consists of m such concatenated embeddings. We flatten this matrix to create a single vector $\tilde{\mathbf{v}} \in \mathbb{R}^{m(d+10)}$ for each article by concatenating all $\tilde{\mathbf{v}}_s$:

$$\tilde{\mathbf{v}} = [\tilde{\mathbf{v}}_1; \tilde{\mathbf{v}}_2; \ldots; \tilde{\mathbf{v}}_m] \tag{5.40}$$

This flattened representation is then processed through a multi-layer feed-forward neural network. The network consists of an input layer, one hidden layers, and an output layer. The input layer takes the flattened vector $\tilde{\mathbf{v}}$, which has a dimension of $m(d+10)$, where m is the number of

[6] In our implementation, we applied normalization to the overall loss. The original FRISS implementation did not describe any normalization of the total loss $J(\theta)$, resulting in loss values significantly exceeding those of the supervised model, thereby disrupting the learning process. Adjusting the loss scaling factor α also failed to mitigate this substantial distortion.

sentences and d is the sentence embedding dimension. Each hidden layer applies a series of operations: a linear transformation, followed by layer normalization, ReLU activation, and dropout. This process can be represented for the first hidden layer as:

$$\mathbf{h}_1 = \text{Dropout}(\text{ReLU}(\text{LayerNorm}(\mathbf{W}_1\tilde{\mathbf{v}}))) \tag{5.41}$$

where $\mathbf{W}_1$ represents the linear layer of the initial layer, $\tilde{\mathbf{v}}$ denotes the sentence embedding, Dropout functions as the dropout layer, ReLU acts as the activation function, and LayerNorm serves to normalize the layer weights.

The layer dimensions reduce step by step. The last output layer only performs a linear transformation without layer normalization, activation functions, or dropout, therefore resulting in a vector $\mathbf{y}_{sent}$ of size C, which denotes the number of frame categories to be predicted. This design gradually reduces the input dimensions from $m(d+10)$ to C, where the final output represents the raw scores for each frame category.

Supervised prediction The averaged representations from both classifiers are then used to predict document-level frames (Step 9 and Step 10). We simply sum the span- and sentence-based prediction:

$$\mathbf{y}_{\text{supervised}} = (\mathbf{y}_{span} + \mathbf{y}_{sent}) \in \mathbb{R}^C \tag{5.42}$$

These predictions are compared to the actual classes using a loss function for classification training (Step 9 and Step 10).

The goal is to balance unsupervised and supervised losses to achieve effective embedding reconstruction and precise frame prediction. This approach entails adding the span-based and sentence-based losses from the supervised segment, adjusting them with the hyperparameter $(1-\alpha)$, and then merging them with the unsupervised loss scaled by α. The total combined loss is:

$$L_{\text{total}} = \alpha \cdot J(\theta) + (1-\alpha) \cdot \left(L_{\text{span}} + L_{\text{sent}}\right) \tag{5.43}$$

where L_{span} and L_{sent} are the losses for the span-based and sentence-based classifiers respectively. This total combined loss is minimized during training to ensure efficient learning.

The SLMuSE-DLF model enhances the FRISS model introduced by [53]. Initially, by integrating Semantic Axis values as an additional perspective for dictionary learning (Step 2b - 5b) and adjusting the supervised component (Steps 7 and 9), the model is able to capture subtle semantics related to framing. As a result, the supervised component is refined to more precisely predict document-level frames by combining latent semantic learning representations.

5.3.4. MuSE-DLF: Multi-Label MuSE-DLF Model

The MuSE-DLF model is an extension of the SLMuSE-DLF model discussed in Section 5.3.3, with adjustments made to facilitate multi-label classification. The primary change includes transitioning from single-label to multi-label classification, allowing the model to predict several categories for each document. This necessitated significant architectural modifications.

Modified Latent Representation Derivation To begin with, the approach for deriving latent representations has been altered. Within the SLMuSE-DLF model, the latent representations $\mathbf{d}_s^z$ and $\mathbf{d}_s^{fx}$ are calculated using a softmax function, which produces a probability distribution across the classes. In contrast, the MuSE-DLF model utilizes a sigmoid function instead of the softmax function. The sigmoid function yields independent probabilities for each class label, thus facilitating multi-label classification. The revised latent representations are formulated as follows:

$$\mathbf{d}_s^z = \sigma(\mathbf{l}_s^z), \quad \mathbf{d}_s^{fx} = \sigma(\mathbf{l}_s^{fx}), \tag{5.44}$$

where σ denotes the sigmoid activation function. This function outputs values ranging from 0 to 1, representing the likelihood of each class being present.

Revised Normalization Technique Furthermore, the normalization technique for latent representations has also been adjusted. The SLMuSE-DLF model employs the Gumbel softmax for normalization. In contrast, the MuSE-DLF model utilizes the Gumbel sigmoid function. This adjustment ensures that the model can produce distinct and sharper class likelihoods, which is particularly necessary for multi-label classification tasks. The normalization process is thus defined as follows:

$$\mathbf{g}_s^z = \text{GumbelSigmoid}(\mathbf{l}_s^z), \quad \mathbf{g}_s^{fx} = \text{GumbelSigmoid}(\mathbf{l}_s^{fx}), \tag{5.45}$$

where GumbelSigmoid represents the normalization function using Gumbel sigmoid.

Preserved Architecture Components With the exception of the stated changes, the architecture and processing pipeline of the MuSE-DLF model remain consistent with the SLMuSE-DLF model. The steps for embedding extraction, concatenation, dimensionality reduction, latent representation restoration, and loss calculation are preserved. The sentence-based classifier described above operates similarly by processing the flattened representation $\tilde{\mathbf{v}}$ using a multilayer feedforward neural network. Instead of employing a softmax function to represent a probability distribution over exclusive categories, we use a sigmoid function. As a result, the final output $\mathbf{y} \in \mathbb{R}^C$ now shows the likelihood of each of the frame categories C occurs within the document.

6.

Experiments

This chapter describes the fine-tuning of the RoBERTa transformer model, followed by the fine-tuning and evaluation of the SLMuSE-DLF and MuSE-DLF models through two separate experiments. Experiment 1 (see Section 6.2) will concentrate on fine-tuning and evaluating the SLMuSE-DLF model, whereas Experiment 2 (see Section 6.3) will target the MuSE-DLF model.

6.1. Transformer Model: RoBERTa

In this thesis, the RoBERTa transformer model was employed[1], chosen for its demonstrated efficiency in frame classification as highlighted by Khanehzar et al. [53] within the FRISS framework. The RoBERTa model provides contextualized embeddings, which are utilized for the computation of Semantic Axis values and for generating word representations for the SLMuSE-DLF and MuSE-DLF models.

6.1.1. Overview of RoBERTa

RoBERTa (**R**obustly **o**ptimized **BERT a**pproach) [62] is an advancement of the BERT transformer model, created in 2019 by Facebook AI Research [62, 64]. This model excels beyond the original BERT in benchmarks like GLUE, RACE, and SQuAD[2] by employing novel pre-training methods [96, 62]. Remarkably, RoBERTa trains on a much larger dataset of 160GB, in contrast to BERT's 16GB, and uses dynamic masking that changes the masked tokens each epoch, thus improving its dependence on contextual information over positional cues [97, 62].

[1] For ease of use and dependability, the fine-tuning was performed using the *FacebookAI/roberta-base*[41] model available on Hugging Face.

[2] GLUE, RACE, and SQuAD are benchmarks in NLP that evaluate models' performance: GLUE tests language understanding in nine different tasks, RACE measures comprehension through English exam questions, and SQuAD examines the ability to respond to questions based on Wikipedia articles and identifies when data is too limited for a reliable response [63].

6.1.2. Fine-Tuning of Transformer Model

The labeled data from the Media Frames Corpus (MFC)[3] [14] was used for fine-tuning in Experiment 1, whereas the SemEval 2023 dataset supplied the labeled data for Experiment 2. Both datasets offer a wealth of domain-specific texts, making them ideal for fine-tuning. The fine-tuning of RoBERTa was performed using masked language modeling (MLM).

During the model optimization stage, perplexity (PPL) served as the main metric to evaluate the performance of the model [42]. Perplexity assesses the alignment between the model's predicted probability distribution and the actual word distribution in the text [42]. A lower perplexity score signifies a more accurate representation of the true distribution, indicating a better grasp of the relevant domain and thereby improving the precision of frame classification in subsequent tasks. The perplexity is calculated by exponentiating the cross-entropy loss (see 6.1) [84]. This metric is particularly beneficial for MLM tasks, as it effectively measures the model's ability to predict contextually suitable tokens [84].

$$\text{PPL} = e^{H(P,Q)} \tag{6.1}$$

Figure 6.1.: *Perplexity (PPL) is calculated by exponentiating the cross entropy between the true distribution P and the predicted distribution Q. Typically, the base* **e** *is used, though other bases like* 2 *or* 10 *can also be used. [44, 84]*

The aim of the fine-tuning process was to lower perplexity, choosing the model with the lowest perplexity for further enhancement. This ensures that our model is finely tuned to the specific domain of our data and more efficient in future classification tasks.

6.2. Experiment 1 (E1): Development of the SLMuSE-DLF model

The primary goal of Experiment 1 is to establish a baseline by implementing the FRISS model from scratch using PyTorch. This experiment seeks to enhance the baseline model by integrating Semantic Axis as an additional component, aiming to improve both the explainability and the performance—specifically in terms of accuracy and macro F1 score—of the model. The integration involves developing the Semantic Axis components independently and then incorporating them into the FRISS model to create a custom model, hereby referred to as SLMuSE-DLF. This new model is designed to leverage the nuanced semantic orientations provided by Semantic Axis together with the semantic role lables to offer more explainable and accurate single label frame prediction.

[3] In reference to the FRISS model [53], the pretrained RoBERTa model was fine-tuned with labeled data to ensure consistency with prior models

6.2.1. Dataset: Media Frames Corpus (MFC)

For the first experiment we utilize the Media Frames Corpus (MFC) by Card et al. [14]. The Media Frames Corpus consists of a collection of news articles annotated with document- and span-level frame labels derived from a set of 15 general frames as defined in *The Policy Frames Codebook* [12], along with the overall tone of the article. This corpus addresses five policy issues: immigration, smoking, gun control, the death penalty, and same-sex marriage. The frames are detailed in Table A.1. The tone can be categorized as anti, neutral, or pro, indicating that articles on the immigration policy issue labeled as anti are against immigration.

The study concentrated on the immigration segment of the MFC, encompassing 6,097 pertinent labeled articles along with 41,286 unlabeled articles. Sourced from 13 prominent US newspapers, these articles span publications from 1971 to 2017. Each labeled article comes annotated with a document-level frame, a primary tone, and other details such as the year of publication, the newspaper of origin, and the headline frame. A detailed analysis of the Media Frames Corpus can be found in the appendix A.1.

6.2.2. Preparation

In the following sections, the preparation required for Experiment 1 is described. Initially, the detailed process of fine-tuning a RoBERTa model on the Media Frames Corpus is explained, followed by the precomputation and evaluation of the semantic axis data.

Fine-tuning of RoBERTa

As described in the methodology section, this thesis employs the RoBERTa pre-trained transformer model[4] for several tasks that require contextualized word embeddings. In the initial experiment, labeled articles from MFC were utilized to fine-tune the RoBERTa model via the Hugging Face *Trainer*, using Perplexity (PPL) as the evaluation metric. Hyper-parameter tuning was conducted through Grid-Search across various batch sizes and learning rates. Batch sizes of 8, 16, 24, 32 and five learning rates ranging from 0.0001 to 0.000001 were tested. The best nine settings, trained over 20 epochs, are displayed in figure 6.2. The most effective configuration used a batch size of *24* and a learning rate of *0.000025*, and was run for 80 epochs on eight Tesla V100-SXM2-32GB GPUs, with an early stopping rule implemented at 15. Figure 6.3 illustrates the PPL learning curve, showing a reduction in perplexity across the global steps, with a low point around approximately 2.6. A model checkpoint near this minimum was preserved for subsequent tasks.

Analysis of Semantic Axis Results

In our initial study, we determine the bias and intensity levels for five selected semantic axes (refer to Section 5.2.2 for selection specifics). The methods described in Section 5.2.3 are employed to

[4]We used the *"FacebookAI/roberta-base"* distribution from Hugging Face [62].

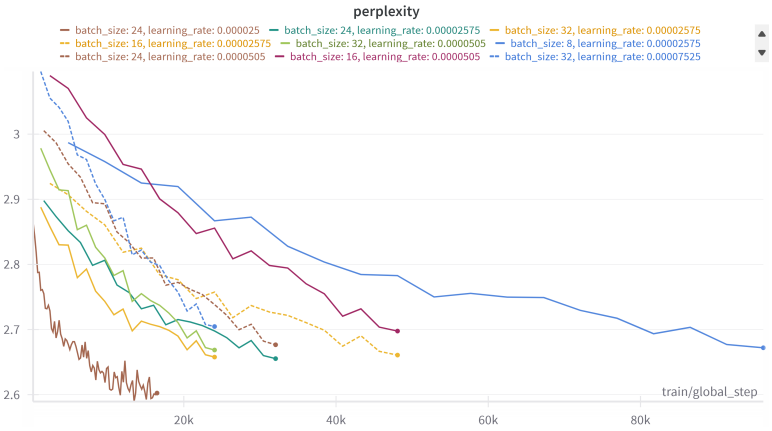

Figure 6.2.: *Perplexity trends throughout hyperparameter tuning. The x-axis denotes global training steps, and the y-axis shows perplexity measurements. Each curve represents a unique pairing of batch size and learning rate, as described by the legend. Various hyperparameter setups were tested in the grid search, with the nine top-performing configurations illustrated here. Lower perplexity signifies superior model performance.*

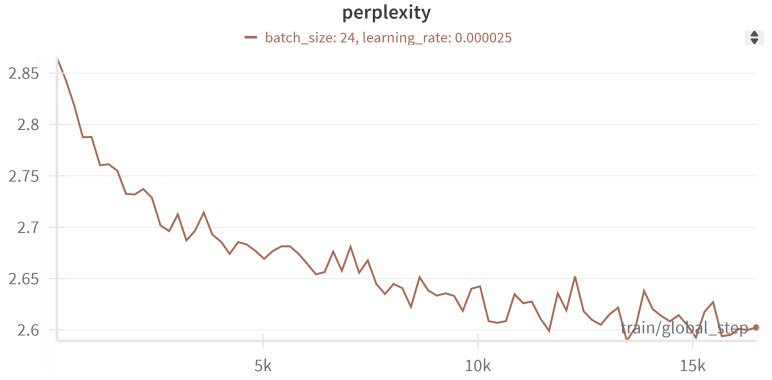

Figure 6.3.: *Progression of perplexity during model training under the best hyperparameter settings. The x-axis denotes the global training steps count, while the y-axis indicates the perplexity score. The graph shows model performance with a batch size of 24 and a learning rate of 0.000025, highlighting a general decrease in perplexity across roughly 15,000 training steps.*

obtain these metrics for each sentence across all articles in the dataset. To enhance the training efficiency of the SLMuSE-DLF model, we have precomputed the bias and intensity values for these semantic axes. This precomputation allows also the evaluation of Semantic Axis values and enriches our understanding of the MFC dataset. With the Media Frames Corpus and annotations on document frames and tones, Semantic Axis enables us to explore the impact of semantic properties on the annotated frames and tones. This section will delve into the analysis of the MFC dataset using the Semantic Axis measurements, focusing on the labeled frames and tones.

To begin analyzing the distribution of semantic axis bias among document frames, we can use a boxplot to display the semantic axis bias values for each document frame. Figure 6.4 shows

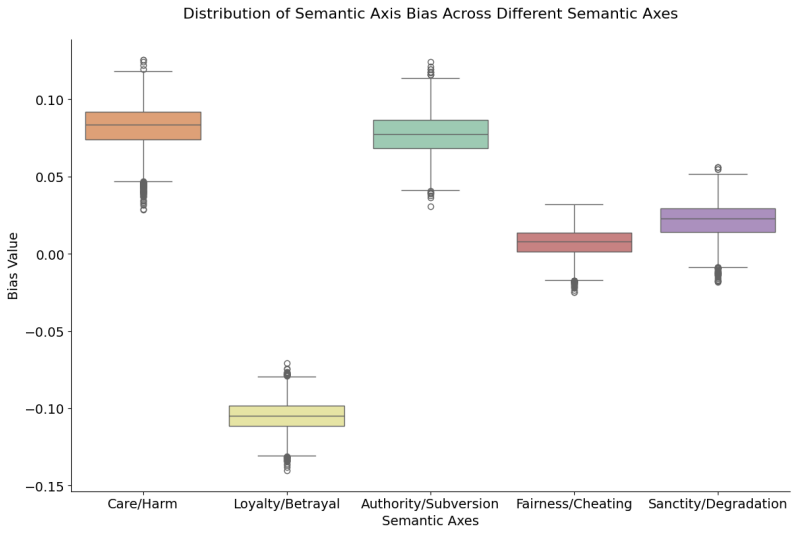

Figure 6.4.: *Distribution of semantic axis bias spanning five moral foundations within the Media Frames Corpus. The boxplots display the extent and central tendencies of bias values for each semantic axis: Care/Harm, Loyalty/Betrayal, Authority/Subversion, Fairness/Cheating, and Sanctity/Degradation. Each boxplot reflects the complete set of bias values computed for its respective semantic axis across all documents in the corpus. Bias values range from -0.15 to 0.15; positive values suggest a preference for the first term in each pair (e.g., Care, Loyalty), while negative values indicate a preference for the second term (e.g., Harm, Betrayal). The plot shows varying levels of bias and dispersion across different moral dimensions.*

that the bias values of the semantic axes are grouped together, indicating a significant separation from the ranges of other semantic axes. The clusters for the *Care/Harm* and *Authority/Subversion* semantic axes are more positively positioned compared to the other semantic axes, which can be interpreted as using more *virtue* or *caring* words for the *Care/Harm* semantic axis and more *authoritarian* words for the *Authority/Subversion* semantic axis. Conversely, the *Loyalty/Betrayal* articles contain a higher frequency of *betrayal* words, which is indicated by the cluster's placement in more negative ranges.

A more detailed analysis of the distribution can be obtained by concentrating on the values for a single semantic axis, specifically the *Care/Harm* semantic axis. Figure 6.5 displays the bias values of the *Care/Harm* semantic axis, categorized by the annotated document-level frame and visualized through a boxplot. This provides insights into how different language is used across various document-level frames. It is evident that this method reveals substantial differences among the frames. For example, the mean of the *Crime and Punishment* document frame is significantly lower than others, and its overall distribution is more negatively skewed compared to other frames. This suggests that articles tagged with the *Crime and Punishment* frame tend to use language that leans towards the *vice* end of the *Care/Harm* semantic axis, effectively using more *harming* words relative to other articles. On the other hand, articles tagged with the *Economic* frame show a higher mean than all other articles, indicating a predominant use of *virtue* words in *Economic* articles

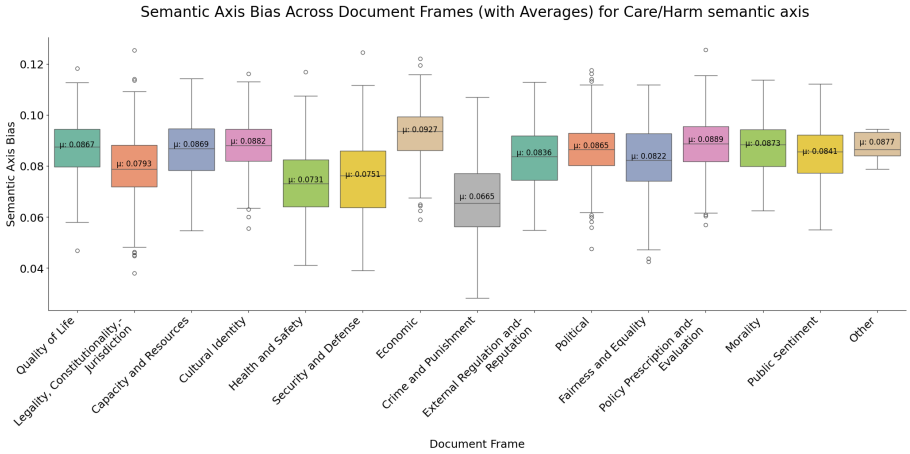

Figure 6.5.: *The figure illustrates the distribution of* Care/Harm *semantic axis bias across various document frames in the Media Frames Corpus. Each boxplot displays the distribution of* Care/Harm *bias values for documents categorized under a particular frame, with mean values (μ) indicated. The y-axis represents semantic axis bias values, theoretically ranging from -1 to 1, where higher values denote a greater emphasis on care-related language. Significant variability in bias distribution is observed among different document frames, highlighting diverse narrative approaches to the Care/Harm dimension. Frames like 'Economic' and 'Cultural Identity' show higher mean bias values, indicating a more care-oriented framing in these contexts.*

compared to those not labeled as such. Referring back to figure 6.4, this pattern is consistent across the other four semantic axes (see Figure B.1 in Appendix B).

Moreover, the study explores the relationship between semantic axis biases and the overall tone of the articles. As explained in Section 6.2.1, an article can be pro-immigration, neutral-immigration, or anti-immigration. For simplicity, the following analysis focuses only on pro-immigration and anti-immigration tones. As shown in Figure 6.6, a pattern emerges where articles with an *anti-immigration* tone exhibit more negative biases across all semantic axes compared to those with a *pro-immigration* tone.

Incorporating semantic axis intensity provides further insights into how a semantic axis is utilized within a specific target corpus. Figure 6.7 demonstrates the relationship between the average bias of each document frame and its average semantic axis intensity for the *Care/Harm* semantic axis. As previously mentioned and shown in Figure 6.5, *Crime and Punishment* exhibits a more negative bias, while *Economic* shows a more positive bias. This pattern is also evident in Figure 6.7, where *Crime and Punishment* appears on the far left of the x-axis, and *Economic* appears on the far right. Notably, *Crime and Punishment* displays the highest semantic axis intensity (0.0023) but a relatively low bias (0.068), indicating a significant use of the *Care/Harm* semantic axis with a slight inclination towards the negative Harm pole in these articles. Conversely, articles labeled as *Other* (indicating a lack of alignment with any of the 14 frames recognized by Boydstun and Gross [12]) display a markedly lower intensity, suggesting a minimal usage of this semantic axis and indicating that articles categorized under the *Other* frame do not exhibit a coherent pattern.

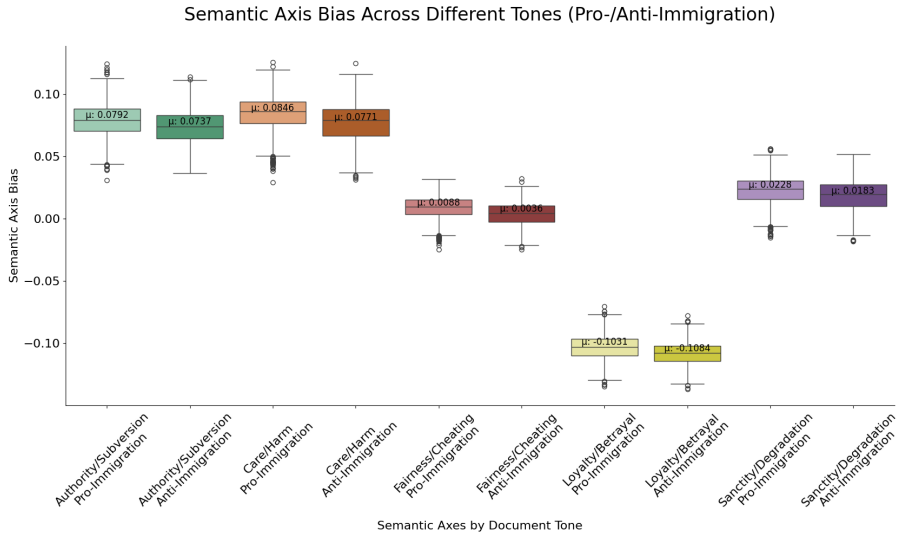

Figure 6.6.: *The distribution of semantic axis bias in pro-immigration and anti-immigration tones within the Media Frames Corpus. The boxplots present the bias values for five semantic axes (Authority/Subversion, Care/Harm, Fairness/Cheating, Loyalty/Betrayal, and Sanctity/Degradation) in both tones. The y-axis indicates semantic axis bias values, where positive values denote alignment with the first term of each pair. Mean bias values (μ) are indicated for each set.*

The *Quality of Life* frame, on the other hand, shows a high bias (0.088) and moderate intensity (0.0021), pointing to language more aligned with the Care pole. This trend is consistent across other semantic axes as well (see Figure B.2 in Appendix B).

By analyzing not only the document frame but also incorporating the primary tone, it is observed that the overall density distribution of all *Crime and Punishment* articles with an *anti-immigration* primary tone (shown in red) tends to lean toward the left side of the x-axis, indicating a generally negative semantic axis bias (see Figure 6.8). In contrast, *pro-immigration* articles (depicted in green) tend to lean towards the right, indicating more positive semantic axis values. This trend is consistent across all five semantic axes and can be seen in Figure B.3 in Appendix B.

Figure 6.9 illustrates how specific terms influence the framing effects by analyzing changes in word-level bias (Eq: 5.22) and intensity (Eq: 5.23). The figure displays the top-5 terms showing the most significant bias shifts within articles discussing the *Legality, Constitutionality, Jurisdiction* frames for the *Fairness/Cheating* semantic axis (see Figure B.4 in Appendix B, view online).

For each dataset (*anti-immigration* and *pro-immigration*), the background corpus includes all articles except those sharing the same frame and tone, serving as a baseline for identifying significant variations in word usage[5]. In the visualization, green bars denote the bias shift in the target corpus, gray bars indicate the bias in the background corpus, and orange bars show the difference between them, which sets the sorting order of the words. Positive orange bars indicate words

[5]This background corpus is essential for revealing distinct patterns and tendencies by comparing the specific tone and topic combinations in the current articles against the complete dataset.

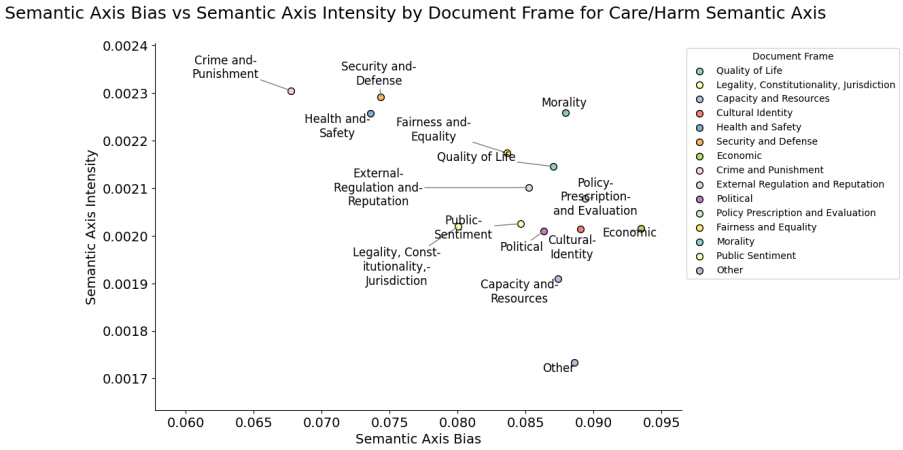

Figure 6.7.: *Correlation between semantic axis bias and intensity for the Care/Harm semantic axis across different document frames. The scatter plot shows the average semantic axis bias (x-axis) and intensity (y-axis) for each document frame within the Care/Harm semantic axis. Semantic axis bias ranges from 0.060 to 0.095, with higher values indicating a greater tendency towards the* virtue *(Care) pole. Semantic axis intensity spans from 0.0017 to 0.0024, indicating how frequently the Care/Harm frame is employed. Notable points include the* Crime and Punishment *frame, which has the highest intensity (0.0023) but a relatively low bias (0.068), signaling frequent use of Care/Harm language with a slight inclination towards the Harm pole. On the other hand, the* Quality of Life *frame shows a high bias (0.088) and moderate intensity (0.0021), pointing to language more aligned with the Care pole.*

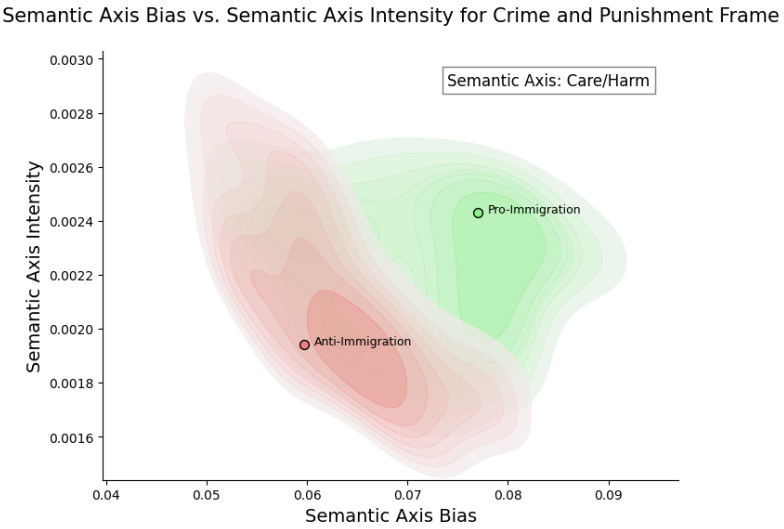

Figure 6.8.: *Scatter plot illustrating the relationship between semantic axis bias (x-axis) and semantic axis intensity (y-axis) for the Crime and Punishment document frame and the Care/Harm semantic axis. The red density plot represents anti-immigration articles, with the average point marked in red, leaning more to the left (indicating lower semantic axis bias values). Conversely, the green density plot represents pro-immigration articles, leaning more to the right, with the average point marked in green.*

that contribute more to the Fairness pole in the target corpus compared to the background, while negative orange bars suggest the opposite.

In *anti-immigration* articles (left subplot), terms like *house* and *state* show smaller word-level bias shifts compared to the background corpus, leaning towards the Cheating pole. Conversely, *Court*, *alien*, and *asylum* tend to move the bias towards the Fairness pole. In *pro-immigration* articles (right subplot), words such as *Right*, *court*, and *asylum* shift the bias more towards the Fairness pole, whereas *House* and *deportation* drive the bias more towards the Cheating pole.

This demonstrates how certain legal expressions distinctly affect the Fairness/Cheating framing in immigration discussions, dependent upon the article's perspective. Notably, some terms such as *court* consistently show a positive bias across both tones, whereas others like *deportation* have varied effects.

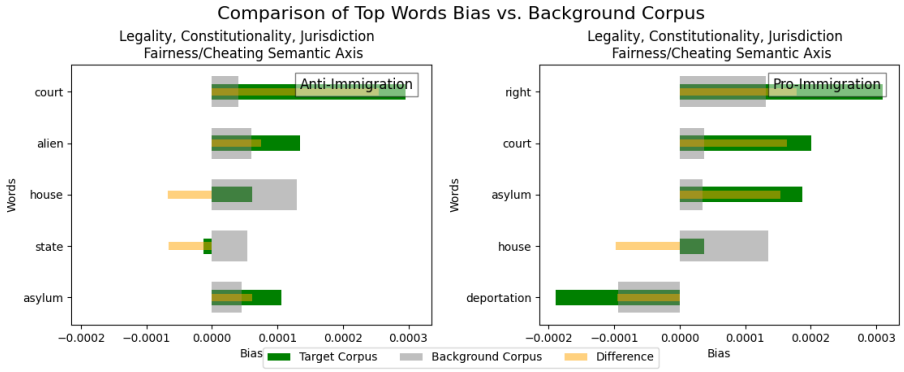

Figure 6.9.: *Analysis of word-level bias within the Legality, Constitutionality, and Jurisdiction frame for the Fairness/Cheating semantic axis, comparing anti-immigration (left) and pro-immigration (right) texts. The graph shows the top five most common words for each corpus and their contribution to the semantic axis bias. Green bars indicate the bias shift in the target corpus (e.g., anti-immigration articles for the left plot), while gray bars indicate the bias in the background corpus (non-target articles within the same frame). Orange bars, which determine the sorting order, represent the difference in bias shift between the target and background corpora. Positive orange bars indicate words that contribute more strongly to the Fairness aspect in the target corpus compared to the background, while negative orange bars indicate the opposite. For example,* court *demonstrates a consistent positive bias across both corpora, while* deportation *shows varying bias between pro- and anti-immigration contexts.*

Figure 6.10 explores the subtleties of framing effects by analyzing the extent of word-level intensity alterations within the *Care/Harm* semantic axis in anti-immigration articles for three distinct document frames. This study demonstrates how certain terms either enhance or weaken the application of specific frames, thereby affecting the overall narrative intensity. For a comprehensive view of intensity shifts across all document frames, tones, and semantic axes, (see Figure B.5 in Appendix B, view online).

Within the *Legality, Constitutionality, Jurisdiction* frame, terms like *court*, *deported*, and *murder* are notably prominent. The frequent appearance of deportation underscores its crucial role in molding the *Care/Harm* narrative in the context of legal immigration discussions. Shifting to the

Security and Defense frame, words such as *attack*, *terrorist*, and *hijacker* are predominant, highlighting the framing of immigration issues from a national security perspective. In the *Economic* frame, terms like *cost*, *employer*, and *billion* are central, illustrating how the *Care/Harm* semantic axis intersects with the economic impact discussions of immigration.

The intensity shifts of these words are logically consistent with their respective frames, demonstrating how language choices can reinforce particular perspectives. For instance, the high intensity of *court* and *deported* in the legal frame naturally aligns with judicial and procedural aspects of immigration, while the prominence of *attack* and *terrorist* in the *Security and Defense* frame vividly illustrates the framing of immigration as a potential threat. Similarly, the focus on *cost* and *billion* in the *Economic* frame emphasizes the financial dimensions of the immigration debate.

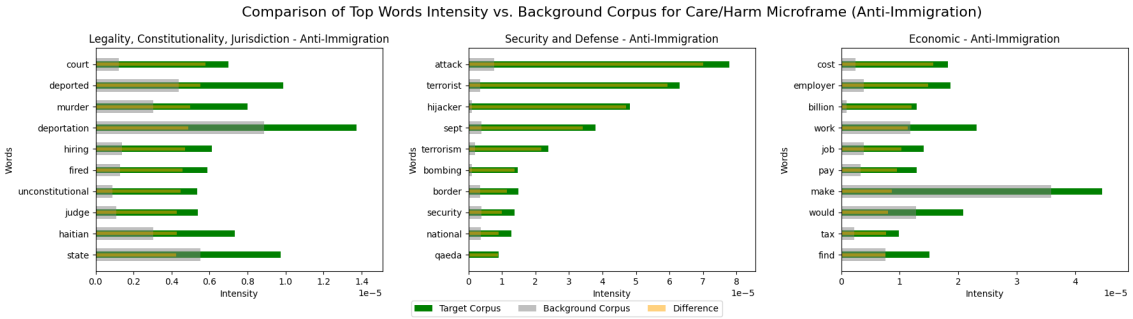

Figure 6.10.: *The analysis of word-level intensity shifts for the Care/Harm semantic axis across three document frames in anti-immigration articles. This figure contrasts the intensity of top words in the target corpus (green) with those in the background corpus (grey) for the Legality, Constitutionality, Jurisdiction (left), Security and Defense (center), and Economic (right) frames. The orange bars indicate the difference in intensity between the target and background corpora, which determines the sorting order. Intensity values on the x-axis are scaled by* 10^{-5}*. This visualization highlights how specific terms influence the Care/Harm framing intensity in different contexts of anti-immigration discourse. Key observations include the high intensity of deportation in the Legality frame, attack in the Security frame, and cost in the Economic frame, suggesting these terms are crucial in shaping the Care/Harm narrative within their respective frames.*

6.2.3. Fine-tuning of SLMuSE-DLF

The upcoming sections outline the rationale behind the constraints imposed on dataset preprocessing and provide a detailed explanation of the fine-tuning procedure for the SLMuSE-DLF model.

Dataset Settings

The dataset constraints, such as the number of sentences or the length of each sentence, were selected to balance data coverage and computational limits. These constraints are detailed in 6.1.

The constraints of these datasets, now referred to as dataset settings, were established through a statistical analysis of the Media Frames Corpus, as depicted in Appendix A in the Tables A.5 and A.6.

Metric	Value
Number of sentences per document	24
Maximum sentence length	64 tokens
Maximum semantic roles per sentence	10
Maximum semantic role length	18 tokens

Table 6.1.: *MFC Dataset Settings*

We exclusively analyze articles from the Media Frames Corpus that are marked as relevant and have a specified document-level frame. This leads to a subset of 5893 articles, which are then examined in more detail. Initial analysis of this subset revealed that limiting sentences to a maximum of 52 words covers around 99% of the corpus sentences. However, our experiments indicated that the word-to-token ratio in our dataset is approximately 1:1.31[6]. To account for this difference, we adjusted the maximum sentence length to 64 tokens, which is the nearest lower multiple of 8 that fits within memory constraints.

As shown in Table A.6, 95% of sentences have 10 or fewer semantic roles, and 95% of semantic roles consist of 14 words or less. By using a 1:1.31 word-to-token ratio, we set the maximum semantic role length to 18 tokens, rounding to the nearest multiple of 8 that complies with memory constraints.

The parameter adjustments offer good coverage of the dataset's properties while remaining within our limited computational constraints. The configurations now cover at least 95% of the dataset's features, including sentence length, number of semantic roles, and role lengths, based on the word-to-token ratio from our empirical analysis.

Fine-Tuning of SLMuSE-DLF

The model's fine-tuning involved a combination of methodical hyperparameter adjustments and practical experiments. Initial parameters, sourced from the FRISS paper by Khanehzar et al. [53], were used as a starting point for further optimization. This tuning process merged a comprehensive approach to testing various configurations with a structured exploration within defined limits.

The initial set of hyperparameters, which encompasses the learning rate, the batch size, the dropout rates, and other parameters essential for the unsupervised module, were sourced from the FRISS paper by Khanehzar et al. [53]. These parameters established the baseline, with subsequent tuning exploring ranges, such as learning rates from 0.00001 to 0.001 and batch sizes from 8 to 64. Initially, Bayesian search methods were used with earlier versions of the architecture. As insights into optimal parameter ranges grew, the focus shifted to more targeted experiments with reliable combinations. The fine-tuned specific hyperparameters included hidden dimensions, dropout probability, learning rate, and various unsupervised component parameters, such as the number of negative samples for contrastive loss and the orthogonality regularization parameter.

[6]We employed a RoBERTa-base tokenizer, resulting in an average of 1.31 tokens per word in our dataset. Refer to Appendix B.1 for more information on the tokenization methodology.

The experiments utilized a setup with four GPUs operating in parallel to speed up the training process. The performance of the model was measured based on *accuracy*, the primary metric used in the baseline FRISS model [53], facilitating comparisons between the SLMuSE-DLF model, the FRISS model and previous models. To avoid overfitting and eliminate redundant computations, early stopping criteria were defined. In particular, training ceased if the accuracy fell below 0.3 after the second epoch due to the unlikely chance of notable progress. Furthermore, a patience-based early stopping method assessed the model four times per epoch and stopped training if no accuracy improvement was observed for more than 20 consecutive training evaluations.

At the end of each epoch, the existing model is evaluated using a test dataset. The complete dataset is divided into training and testing sets by applying stratification to maintain the same class distribution in both subsets. We use a test size of 10%, thus ensuring that 90% of the data remain for training, which is sufficiently large, and 10% is reserved for testing, providing a large enough evaluation set for the model.

Initially, CrossEntropyLoss was utilized as the loss function; however, to address the imbalance of the dataset, the focal loss with class-specific alpha weighting was subsequently used. Focal Loss, as introduced by [61], is especially effective for classification tasks with severe class imbalances, as it reduces the influence of easy examples and concentrates on hard-to-classify instances. The detailed implementation and mathematical formulation of Focal Loss used in this work are presented in the Appendix B.2.1. In our PyTorch implementation, we utilize the PyTorch implementation from Hassan [36].

In our implementation of Focal Loss, γ is set to 2.0, and α^{fl} is derived through a multistep process to address class imbalance. α^{fl} should not be mistaken for the α parameter used for balancing supervised and unsupervised loss. first establish a minimum frequency threshold of 5% and adjust any class frequencies below this threshold. The adjusted frequencies are then normalized and inverted to give more weight to underrepresented classes. The final α^{fl} values are calculated as:

$$\alpha_i^{fl} = \frac{(1/f'i)}{\sum_j (1/f'_j)} \cdot 100, \tag{6.2}$$

where f'_i represents the adjusted and normalized frequency for class i. This approach effectively balances the contribution of each class to the overall loss while maintaining an emphasis on underrepresented classes, addressing the challenges posed by significant class imbalances in the dataset.

A learning rate scheduler was used. Initially, a linear warm-up scheduler was applied, increasing the learning rate from zero to the set learning rate over the defined warmup steps, and subsequently decreasing it linearly to zero over the remaining batches. However, to avoid local minima, the strategy was switched to CosineAnnealingWarmRestarts. The CosineAnnealingWarmRestarts scheduler adjusts the learning rate using a cosine annealing pattern and periodically resets it to its initial value. This method helps in escaping local minima and encourages better exploration of the loss landscape. The scheduler was set up with an initial cycle $\mathbf{T_0}$ length that is either one-third of the total training steps or one epoch, whichever is greater. Each following cycle doubles in duration ($\mathbf{T}_{mult} = 2$). The minimum learning rate eta_min was fixed at 1% of the starting learning rate.

This setup allows for adaptive learning rate adjustments throughout training, potentially enhancing convergence and generalization performance.

To maintain computational efficiency while ensuring sufficient training, the number of epochs was capped at 15.

Table 6.2 displays the hyperparameter ranges that were evaluated and the ultimate settings for the optimal model.

Parameter	**Range Explored**	**Final Value**
Learning Parameters		
Learning Rate	0.00001 - 0.001	0.00005
Batch Size	8 - 64	64
Dropout Probability	0.1 - 0.4	0.3
Optimizer	adam, adamw	adamw
Adam/AdamW Weight Decay	0.000001 - 0.001	0.00001
Adam/AdamW use AMS Grad	True, False	True
Model Dimensions		
Hidden Dimension	768 - 2048	768
Number of Layers (Unsupervised)	1 - 3	3
Number of Layers (Unsupervised FrameAxis)	1 - 3	3
Activation and Normalization		
Activation Function (Unsupervised)	relu, gelu	gelu
Use Layer Norm (Unsupervised)	True, False	True
Activation Function (Unsupervised FrameAxis)	relu, gelu	gelu
Use Layer Norm (Unsupervised FrameAxis)	True, False	True
Pooling and Embedding		
Sentence Pooling	mean, cls	mean
Hidden State for Embedding	last, second-to-last	second-to-last
Matmul Input (Unsupervised)	g, d	g
Matmul Input (Unsupervised FrameAxis)	g, d	g
Gumbel Softmax Log (Unsupervised)	True, False	False
Gumbel Softmax Log (Unsupervised FrameAxis)	True, False	False
Loss and Regularization		
Alpha (α)	0.1 - 0.6	0.5
Lambda Orthogonality (λ)	0.001 - 0.1	0.001
Max Margin for Focal Triplet Loss (M)	8	8
Number of Descriptors (t)	8	8
Number of Negative Samples ($N\_$)	32 - 128	128
Focal Loss Gamma (γ)	1-4	2
Temperature and Decay		
Tau Decay (τ)	0.0001 - 0.01	0.0005
Minimum Tau (τ_{min})	0.5	0.5

Table 6.2.: *Summary of the Final Hyperparameter Configurations and Assessed Ranges*

6.2.4. Results and Evaluation

The following sections present the gathered model metrics, interpret and clarify these metrics, and assess the explainability of the SLMuSE-DLF model.

Metrics

Our SLMuSE-DLF's performance was assessed using the accuracy and macro-F1 metric, aligning with the metrics utilized in earlier frame prediction models based on the Media Frames Corpus. Table 6.3 presents a comparison of SLMuSE-DLF against other models trained on the MFC.

Model	Acc.	Macro-F1
FRISS [53]	0.697	0.605
Khanehzar, Turpin, and Mikolajczak [52]	0.658	-
SLMuSE-DLF (Best)	0.643	0.520
SLMuSE-DLF (Mean)	0.606	0.497
Ji and Smith [47]	0.584	-
Field et al. [26]	0.573	-
Card et al. [13]	0.568	-

Table 6.3.: *Performance analysis of various frame prediction models, trained using the Media Frames Corpus*

Our best SLMuSE-DLF model achieved an accuracy of 0.643 and a macro-F1 score of 0.520. While these metrics do not surpass the FRISS baseline, they demonstrate competitive performance in the field of frame prediction. To confirm the reliability of our methodology, we performed a 10-fold cross-validation. The mean accuracy for these folds was 0.606 with a standard deviation of 0.0223, and the macro-F1 score averaged 0.497 with a standard deviation of 0.0259, reflecting different performance levels across various data partitions.

Figure 6.11 illustrates the SLMuSE-DLF learning curves over a span of 14 epochs. The accuracy, represented in solid blue, consistently surpasses the macro-F1 score, depicted by a dashed green line, with accuracy values ranging between 0.4 and 0.62, and macro-F1 values between 0.2 and 0.52. Both metrics display an upward trend interrupted with fluctuations. This difference between accuracy and macro F1 might indicate an improvement in overall correctness. However, it also suggests class-specific[7] variations, possibly due to class imbalance; meaning some classes are predicted accurately, while others face considerable difficulties.

Figure 6.12 illustrates the training loss curve. This curve shows a steady reduction in loss over the course of training, signifying efficient model learning. The sharp initial drop in loss parallels the quick enhancement observed in the accuracy and F1 score curves. The later more gradual decline indicates that the model keeps fine-tuning its predictions, though more slowly.

To gain a more nuanced understanding of SLMuSE-DLF's performance, we examined its F1 scores and precision for individual frames. Figure 6.13 presents the class-specific F1 scores:

The class-specific F1 scores reveal varying performance across different frames. Notably, the model performs exceptionally well on the *Political* frame (F1: 0.810), followed by *Crime and Punishment* (F1: 0.750) and *External Regulation and Reputation* (F1: 0.740). Conversely, the model struggles with the *Fairness and Equality* frame (F1: 0.250) and the *Other* category (F1: 0.000).

[7] In this context, the term class refers to frames. Therefore, the words frame and class can be used interchangeably.

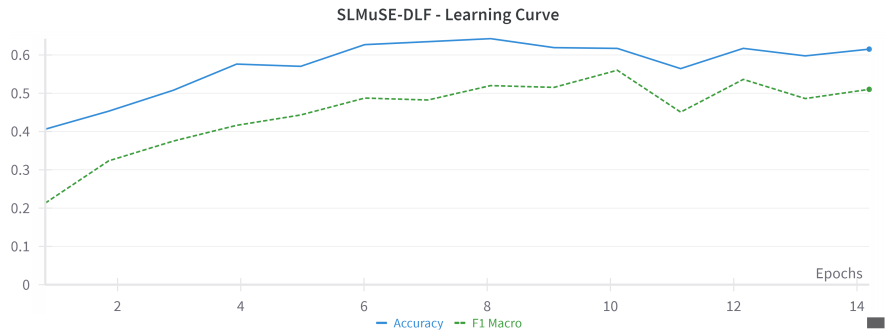

Figure 6.11.: *Performance curve for the SLMuSE-DLF model, illustrating Accuracy and F1 Macro metrics over 14 epochs. The chart demonstrates a general increase for both metrics, with Accuracy (blue line) consistently exceeding F1 Macro (green dashed line). Accuracy reaches its maximum at around 0.65, while F1 Macro attains its highest value close to 0.55, reflecting a continuous enhancement in model performance throughout the training phase.*

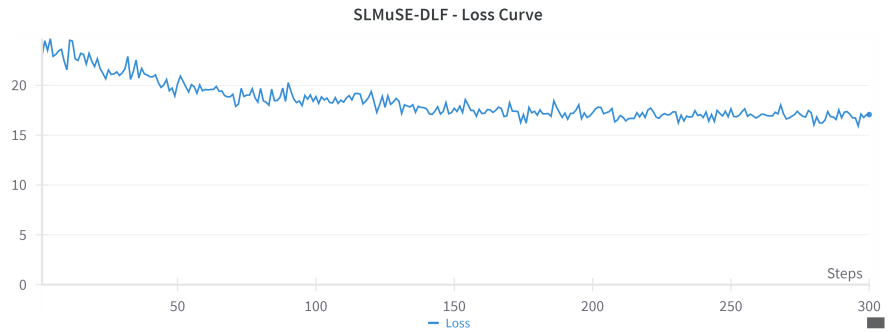

Figure 6.12.: *Training loss curve of the SLMuSE-DLF model across 300 steps. The diagram demonstrates a typical downward trajectory, beginning with a high loss approximately at 25 and leveling off near 17 towards the conclusion of the training. The curve features a swift initial reduction in loss followed by slower progress, indicating successful learning and model convergence over the training duration.*

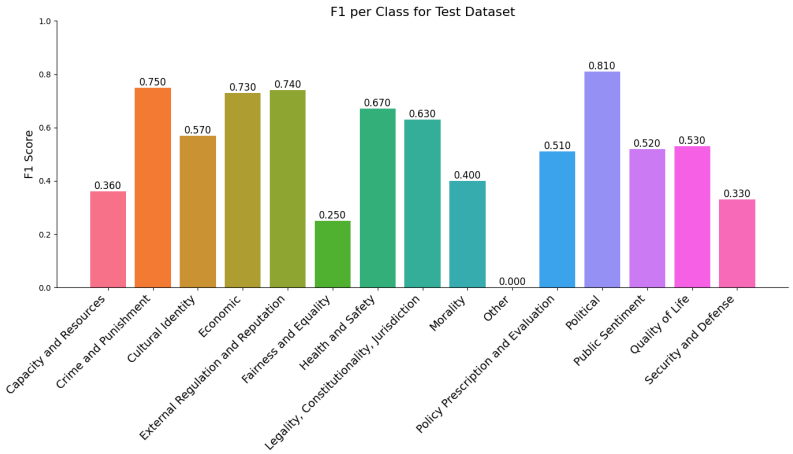

Figure 6.13.: *F1 scores grouped by frame type for the SLMuSE-DLF model evaluated on the test data. Taller bars represent higher scores, indicating an improved balance of precision and recall. Observe the differences among the frames, with 'Political' achieving the highest F1 score of 0.810, while 'Other' has the lowest at 0.000.*

Figure 6.14 shows the precision scores for each class:

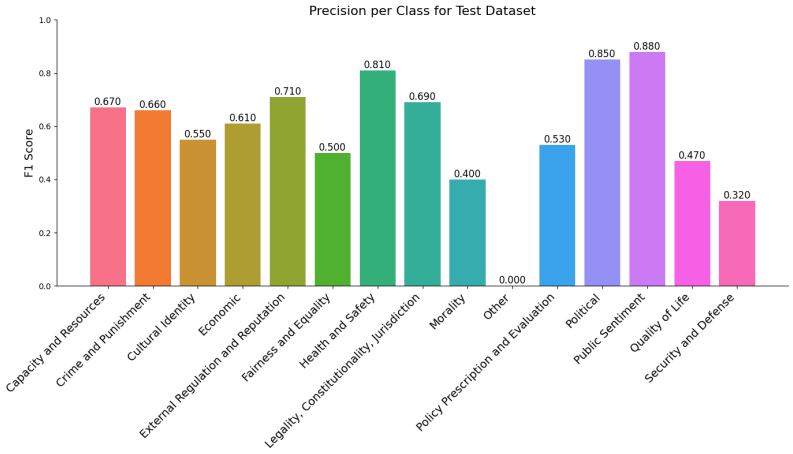

Figure 6.14.: *Precision scores for various frame categories using the SLMuSE-DLF model on the test dataset. Taller bars (higher scores) represent better precision in positive predictions for each frame. Observe the differences among categories, with 'Public Sentiment' having the highest precision (0.880) and 'Other' the lowest (0.000).*

The precision scores shed more light on the model's performance. *Public Sentiment* boasts the highest precision (0.880), with *Political* (0.850) and *Health and Safety* (0.810) closely following. This implies that the model is quite reliable when it predicts these categories. On the other hand, the lower precision scores for *Security and Defense* (0.320) and *Fairness and Equality* (0.500) indicate a tendency for the model to produce false positives in these areas.

Investigating the gaps between F1 scores and precision for each class provides valuable insights into the model's performance. For classes where precision is notably higher than the F1 score (e.g.,

Public Sentiment with 0.880 precision vs. 0.520 F1), it indicates that the model has high precision but lower recall. This means that the model is very accurate in predicting this frame(low number of false positives), but it may miss many instances of it (high number of false negatives). Classes where the F1 and precision scores are similar (e.g., *Political* with 0.810 F1 and 0.850 precision) display a good balance between precision and recall, suggesting strong overall performance for these frames. For classes such as *Capacity and Resources* where the F1 score (0.360) is much lower than the precision (0.670), it shows that the model suffers more from false negatives than false positives. The model rarely incorrectly labels other frames as this one, but it frequently fails to identify this frame when it's present. The *Fairness and Equality* frame with low F1 (0.250) and precision (0.500) scores illustrates that the model struggles both with correctly identifying this frame and avoiding false positives. These differences underscore the necessity of considering both metrics to fully understand the model's strengths and weaknesses across different frames. The model faces particular difficulty with the *Other* frame, as it serves as a fallback category for articles that do not match any other class. Consequently, this leads to a dataset with highly heterogeneous data. Additionally, the *Other* class appears in just 0.15% of all articles.

To better understand these results, it is essential to examine the frame distribution in our dataset, which is detailed in Table A.2 in the Appendix A. The class distribution indicates significant imbalances, likely affecting the model's performance across various frames. The model excels in the *Political* frame, given its substantial presence in the dataset (15.79%). Likewise, the high performance on *Crime and Punishment* (13.02% of the dataset) implies that the model gains from numerous training examples for these prevalent frames.

It is interesting to note that although *Legality, Constitutionality, Jurisdiction* is the second most frequent frame (15.66%), it has a lower F1 score (0.630) and precision (0.690) compared to some less frequent frames. This suggests that this frame might be especially difficult to classify, possibly because of overlap with other legal or political frames.

The model's struggle with the *Fairness and Equality* frame may be partially attributed to its relative scarcity in the dataset (2.54%). However, the high precision for *Public Sentiment* despite its low representation (3.99%) suggests that some minority classes can still be accurately predicted when they have distinct features. In this case, weighting the classes inversely proportional to their frequency could lead the model to heavily penalize incorrect classifications of this class, thereby compelling the model to improve its learning process.

The weak results in the *Other* category are expected due to its very low occurrence (0.15%) and probable diverse characteristics.

Explainability

The SLMuSE-DLF model utilizes various post-hoc explainability methods that adhere to the Explainable AI (XAI) guidelines mentioned in Section 2.2. These methods aim to provide deeper insight into the factors influencing the model's decision-making process and the final frame prediction, thereby enhancing the prediction's explainability. We propose three strategies to render

the MuSE-DLF model transparent. We perform semantic role analysis at both the dataset and article levels, following the work of Khanehzar et al. [53], and we also introduce a novel approach to evaluate the semantic orientation of the articles using data from the semantic axis.

Dataset-level Semantic Role Analysis To gain insights into the most influential words for each frame throughout the entire dataset, we performed a semantic role analysis at the dataset level. We utilize the values $\mathbf{g}^z$ for each span, where $z \in \{agent, theme, predicate\}$. Adhering to the methodology established by Khanehzar et al. [53], a span is assigned to a frame if its threshold exceeds 0.8. Next, we compute the inverse frequency of each word, remove stop words along with other irrelevant words, and assemble a table of the most significant words for each frame and semantic role.

Table 6.4 provides a selection from this analysis, listing the top words corresponding to each semantic role (predicate, agent, and theme) for three specific frames: *Economic*, *Political*, and *Morality*. A complete table for all frames can be found in Appendix B Figure B.2.3.

Frame	**Predicate**	**Agent**	**Theme**
Economic	hire, pay, work, fill, estimated	bank, grower, employer, company, worker	labor, worker, tuition, economic, job
Political	build, decided, approved, sponsored, support	fox, party, republican, romney, senate	politician, ad, reform, pressure, democratic
Morality	need, failed, include, let, remain	trump, official, woman, congress, state	detail, whose, along, housing, cause

Table 6.4.: *Top 5 words identified for the frames Economic, Political, and Morality within the three semantic roles: predicate, agent, and theme.*

This method uncovers trends in how words are utilized across various frames and semantic roles. The *Economic* frame distinctly emphasizes employment and finance. The predicates encompass verbs such as *hire*, *pay*, and *work*, with agents labeled by terms like *bank*, *grower*, and *employer*. The themes in this frame are consistent, including words like *labor*, *worker*, and *economic*.

Similarly, the *Political* frame shows a strong alignment with the political discourse. Predicates such as *build*, *decided*, and *approved* reflect political actions. Agents include political entities and media (*fox*, *party*, *republican*), while themes encompass relevant concepts like *politician*, *ad*, and *reform*.

On the other hand, the *Morality* frame seems less unified. Although certain predicates (*need*, *failed*) may pertain to moral judgments, others appear less directly related. The agents are a combination of politicians and generic terms, and the themes (*detail*, *whose*, *housing*) do not clearly embody moral concepts. This indicates that the model might struggle with consistently identifying or representing the *Morality* frame.

At the dataset level, this semantic role analysis provides valuable understanding of the usage of words with different frames. It enables us to identify the typical agents, themes, and actions related to each frame, assisting us to better understand the model predictions. The analysis also highlights

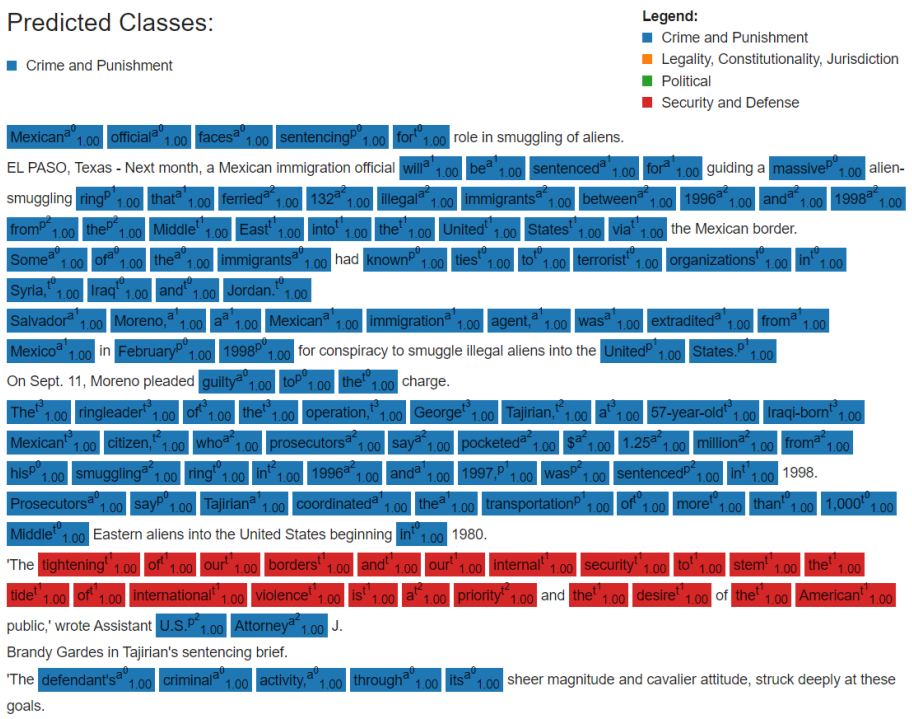

Figure 6.15.: *Illustrations of semantic roles at the article level. The figure illustrates the predicted frame for each semantic role, marked by color. The color intensity and subscript digits reveal the prediction probability, while the superscript denotes the role associated with the semantic role (with p for predicate, a for agent, and t for theme). Linked semantic roles are indicated by the superscript number above the semantic role letter.*

areas where the model performs well (such as *Economic* and *Political* frames) and where it may need improvement (as in the case of the *Morality* frame).

Article-level Semantic Role Visualization This method visualizes frame predictions for sentences within individual articles. Unlike corpus-level techniques that find the words most associated with specific frames across an entire dataset, this method provides clearer insights for single articles. For each sentence, the model examines and constructs a tensor $\mathbf{d}^z$ for each span, where $z \in \{predicate, agent, theme\}$. Each element in this tensor shows the probability that the span predicted a particular frame. A span is categorized into frame types when the probability is at least a threshold value. We set the threshold value at 0.5.

Figure 6.15 demonstrates how effective this visualization method is. In this example, the model accurately predicted the frame *Crime and punishment*, which matches the dominant *Crime and punishment* frames within the spans. The visualization indicates that the majority of the text is highlighted in blue, representing the *Crime and punishment* frame. This showcases the model's precision in recognizing the main frame of the article through span-level predictions.

Figure 6.16 reveals one of the model's weaknesses. In this case, the model wrongly classified the frame as *Political*, whereas the accurate frame is *External Regulation and Reputation*. The

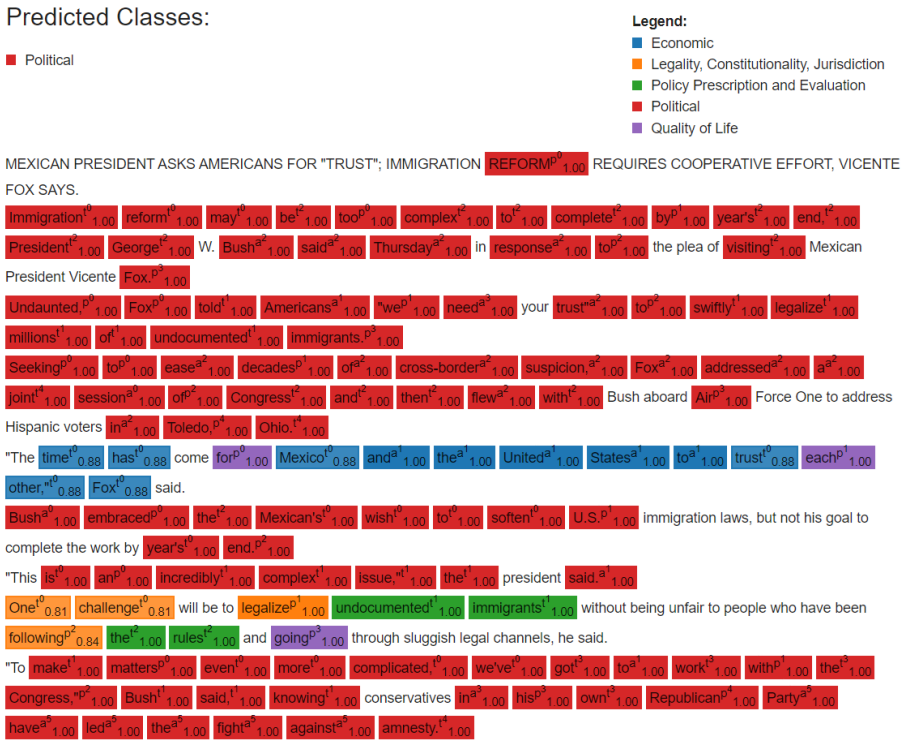

Figure 6.16.: *Depiction of the model's prediction errors. The figure illustrates numerous spans labeled as Political, leading to an inaccurate high-probability prediction for this category across the entire document.*

visualization depicts numerous spans marked as *Political*, resulting in a high-probability prediction for this category at the document level. This instance underscores the necessity for meticulous evaluation of the model's predictions and the potential for errors stemming from span-level categorizations. The prevalence of *Political* spans led the model to an incorrect overall frame prediction.

The visualization method at the article level demonstrates that the span view offers further insight into how the model forecasts the overall frame, helping users grasp the foundation of the model's decisions. However, it is crucial to recognize that the model is not perfect, as evidenced by its performance metrics and instances of misclassification.

Semantic Axis Bias Visualization To improve our comprehension of frame predictions and semantic orientations in articles, we utilized a visualization technique grounded in the Semantic Axis values, which is initially presented in this research. This technique provides a detailed perspective on the alignment of each sentence and the entire article with specific semantic axes. We created these visualizations using the values $\mathbf{d}^{fx}$ with a set threshold of 0.5, assigning semantic axis values to a frame when probability exceeded this threshold. These assignments were then used to determine the background bias to which we compare the semantic inclinations of the article.

Figure 6.17 depicts the evaluation of an article employing the semantic axis method, with an emphasis on the Care/Harm dimension. The model correctly predicted the frame as *Legality, Constitutionality, Jurisdiction* for this article. The overall article bias indicates a tendency towards more harming language. Interestingly, sentences 1 and 3 are categorized as *Crime and punishment*, which is incorrect, while sentence 2 is correctly predicted as *Legality, Constitutionality, Jurisdiction*. This discrepancy highlights the complexity of frame prediction at different levels of granularity. Sentence 1, for example, uses more caring language with words like *APPEAL*, *HOPES*, and *STAY* marked as positive (green box), while *EL SALVADOR* is marked as negative (red box).

The visualization's title reveals the overall semantic tendency of the article. The green bar displays the semantic bias of the entire article, while the gray bar indicates the bias from the background corpus, which is generated using predictions from our model. For instance, when reviewing an article tagged with an *Economic* frame, the background corpus consists of all articles not classified as *Economic*. This approach allows for a more nuanced comparison between the focal article and the larger data collection. The difference between the background (gray) and the current article (green) is highlighted in orange. A negative distinction, indicated by the red arrow pointing downwards next to 'Overall Article Bias,' suggests that the article leans towards the negative side of the semantic axis (Harm, in this case), reflecting a higher prevalence of harming terms as opposed to caring terms.

The graphical representation broadens this analysis to each individual sentence, facilitating a detailed evaluation of how specific words influence the overall semantic inclination. This method allows us to grasp the semantic orientation of articles through their language, uncovering subtleties that might be overlooked in more general classifications.

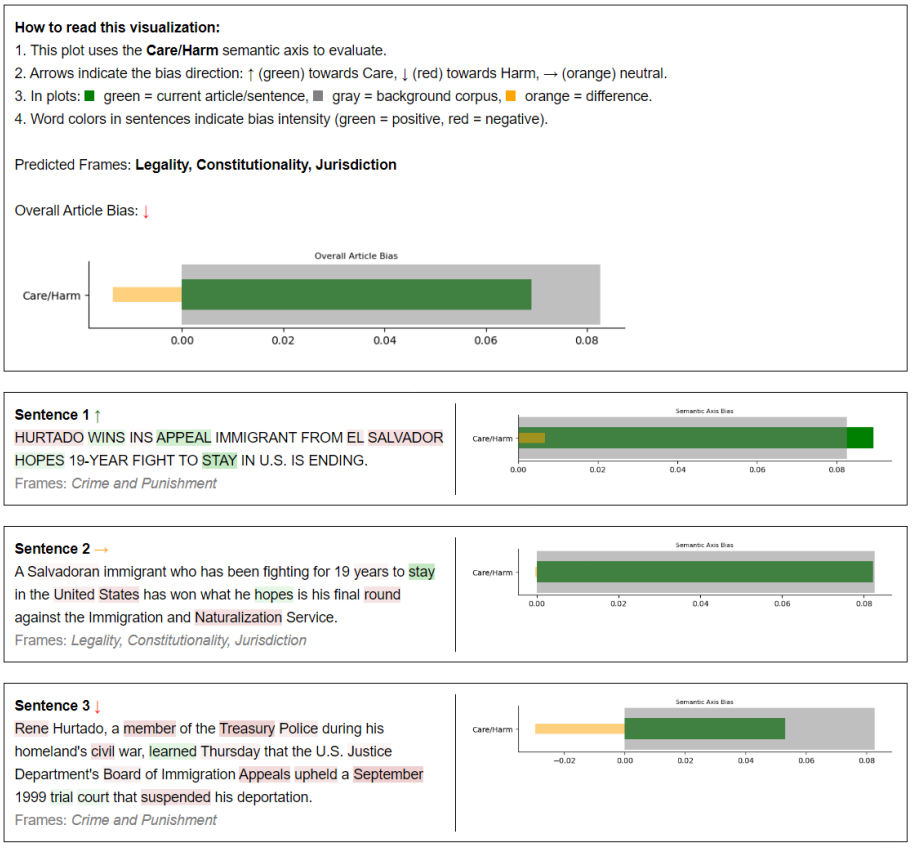

Figure 6.17.: *Depiction of an article's semantic bias along the Care/Harm dimension. The top part shows the overall bias of the article: green represents the current article, gray denotes the background corpus, and orange highlights the difference. Below, individual sentence analyses are displayed, with words color-coded to indicate the bias level (green for positive and red for negative).*

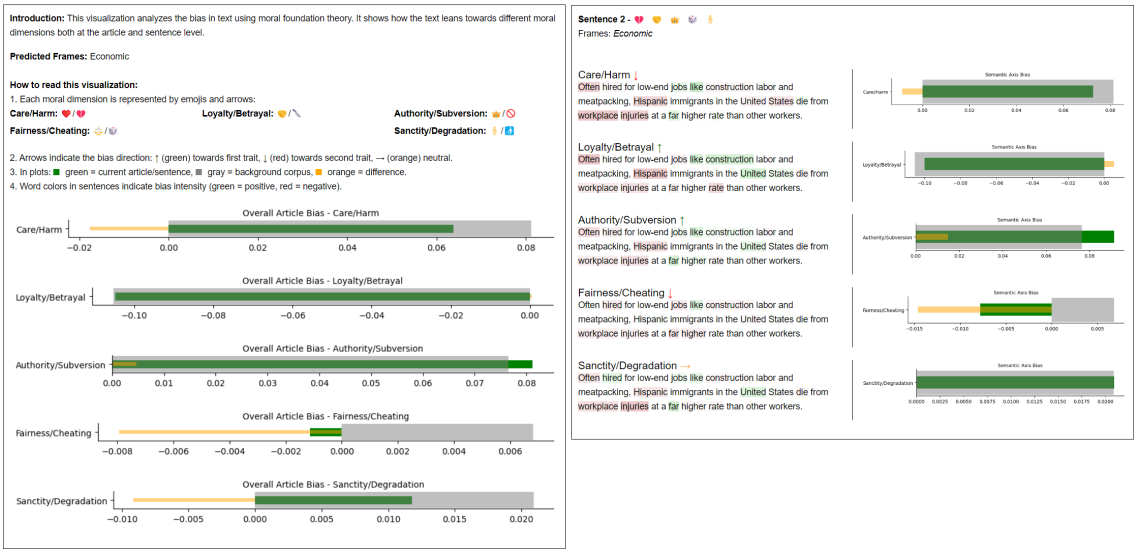

Figure 6.18.: *Assessment of semantic axis bias in an article across five moral dimensions. Each subplot represents the bias for a given semantic axis, showcasing the total article bias at the top and a sentence-wise analysis below. This visualization displays the article along with its associated predicted frames.*

Figure 6.18 broadens this analysis to include all five semantic dimensions for a single text. In this case, the model wrongly predicted the frame as *Economic*, whereas the correct frame is *Fairness and Equality*. This error highlights the difficulties in accurately identifying the main frame of an article.

The visualization offers an extensive overview of the article's semantic biases across the five moral dimensions: *Care/Harm, Loyalty/Betrayal, Authority/Subversion, Fairness/Cheating, and Sanctity/Degradation*. In general, the article is written with a more negative tone, inclining towards harm, betrayal, subversion, cheating, and degradation in these areas.

Reviewing the analysis at the sentence level, it is evident that the initial sentence corresponds with the erroneous document-level frame prediction. This sentence is identified as more harmful, having terms such as *Often*, *Hispanic*, *workplace*, and *injuries* that contribute to the harmful bias. This case demonstrates how single sentences can affect the overall frame prediction, potentially causing misclassifications.

The illustration of semantic axis bias offers several key benefits. It provides understanding at both the article and sentence levels, facilitating a comprehensive understanding of semantic structures. By integrating background corpus data, it places the article's semantic inclinations in the context of a broader dataset. The use of color-coding for individual terms helps identify specific language choices that affect the overall semantic orientation.

6.3. Experiment 2 (E2): Development of the MuSE-DLF

The main objective of Experiment 2 is to enhance the SLMuSE-DLF model from Experiment 1 to predict multiple frames. This experiment aims to make the SLMuSE-DLF model capable of handling multiple labels, utilizing the SemEval 2023 dataset. The goal is to boost both explainability and performance, especially regarding F1-micro and F1-macro scores. These modifications lead to the development of a new model called MuSE-DLF. This new model is crafted to take advantage of the detailed semantic orientations provided by FrameAxis, along with the semantic role labels, to offer more explainable and precise multi-label frame predictions.

6.3.1. Dataset: SemEval-2023 Dataset

The *SemEval-2023 Task 2 Dataset* [78], hereafter called the SemEval dataset, comprises news articles in nine different languages, capturing the major global events from 2020 to mid-2022. Released as part of the SemEval-2023 challenge, this collection covers a wide range of topics, such as the COVID-19 pandemic, abortion laws, migration, the Russo-Ukrainian conflict, and various local events. It includes development, training, and testing data for English, French, German, Italian, Polish, and Russian, as well as testing data for Spanish, Greek, and Georgian (termed surprise languages). The development and training data can be utilized for model training, while the test data is exclusively for validating model performance using a provided validator script. This script compares the predictions on the test dataset with the gold labels and returns the results in terms of micro- and macro-F1 scores, which can be compared to other models. The dataset is annotated using the frames from *The Policy Frames Codebook* [12] and is classified into three types of articles: *opinion*, *reporting*, and *satire*.

An in-depth examination of the SemEval-2023 dataset is provided in Appendix A.2.

6.3.2. Preparation

The following sections outline the preparations necessary for Experiment 2. First, the process of fine-tuning a RoBERTa model on the SemEval dataset is detailed, and then the steps for the pre-computation and assessment of semantic axis data are described.

Fine-tuning of RoBERTa

As described in the methodology section, this thesis employs the RoBERTa pre-trained transformer model[8] for several tasks that require contextualized word embeddings. In the initial experiment, train articles from the SemEval 2023 dataset were utilized to fine-tune the RoBERTa model

[8]We used the *"FacebookAI/roberta-base"* distribution from Hugging Face [62].

via the Hugging Face *Trainer*, using Perplexity (PPL) as the evaluation metric. We follow the fine-tunining setup as utilized in Experiment 1 (see section 6.2.2). We also apply hyper-parameter tuning using Grid-Search across various batch sizes and learning rates. Batch sizes of 8, 16, 24, 32 and five learning rates ranging from 0.0001 to 0.000001 were tested. The best nine settings, trained over 20 epochs, are shown in figure 6.19. The optimal setup involved a batch size of *32* and a learning rate of *0.00002575*. This configuration was executed for 20 epochs using an NVIDIA A100-SXM4-80GB GPU. An early stopping mechanism was in place, which terminated the training if no perplexity improvement occurred over 15 consecutive iterations. Figure 6.20 illustrates the PPL learning curve, showing a reduction in perplexity across the global steps, with a low point on the x-axis around approximately 3.8. A model checkpoint near this minimum was preserved for subsequent tasks.

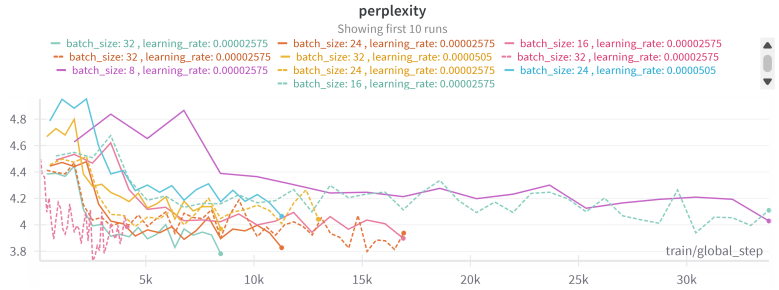

Figure 6.19.: *Perplexity patterns during the hyperparameter tuning of language model training. The chart depicts the initial 10 experiments with different batch sizes (8, 16, 24, 32) and learning rates (0.00002575, 0.0000505). The x-axis indicates global training steps up to 30,000, and the y-axis presents perplexity values ranging from 3.8 to 4.8. Each unique line color and style represents a distinct hyperparameter setup. Lower perplexity values denote improved model performance, with most setups stabilizing between 4.0 and 4.2 after 15,000 steps, indicating relative consistency across the various hyperparameter settings.*

Analysis of Semantic Axis Results

In our initial research, we identified the bias and intensity levels for five chosen semantic axes (see section 5.2.2 for details on selection). The techniques outlined in section 5.2.3 are used to derive these metrics for each sentence in all articles within the dataset, following the same approach as the Media Frames Corpus analysis. Consequently, we also precomputed the bias and intensity values of the semantic axis for the MuSE-DLF model to improve the efficiency of training. This precomputation facilitates the evaluation of Semantic Axis values and enhances our understanding of the SemEval dataset. With the SemEval and annotations on document frames and article types, Semantic Axis allows us to investigate the influence of semantic properties on the annotated frames and article types. This section will explore the analysis of the SemEval dataset using Semantic Axis measurements, concentrating on frames and article types.

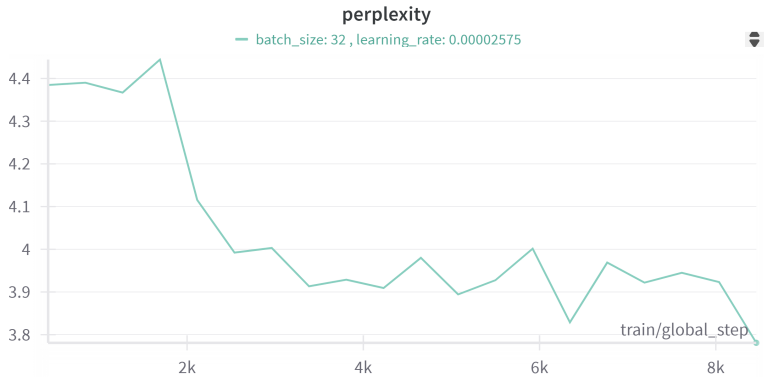

Figure 6.20.: *Perplexity trends for the best hyperparameter setting (batch size: 32, learning rate: 0.00002575) across 8000 training steps. The x-axis denotes training steps, and the y-axis indicates perplexity values from 3.8 to 4.5. The chart shows a sharp initial fall in perplexity up to around 2000 steps, followed by a more moderate decrease and leveling off near 3.9-4.0, signifying enhanced model performance and convergence during training.*

To begin analyzing the distribution of semantic axis bias within the provided articles, we can use a boxplot to display the semantic axis bias values for each semantic axis across all articles. Figure 6.21 shows that the bias values of the semantic axes for the semantic axes *Care/Harm*, *Loyalty/Betrayal*, *Authority/Subversion*, *Fairness/Cheating*, and *Sanctity/Degradation* cluster together, showing significant separation from the ranges of other semantic axes. The clusters for the *Care/Harm* and *Authority/Subversion* semantic axes are more positively positioned relative to the other semantic axes, which can be interpreted as using more *virtue* or *caring* words for the *Care/Harm* semantic axis and more *authoritarian* words for the *Authority/Subversion* semantic axis. In contrast, the *Loyalty/Betrayal* articles contain a higher frequency of *betrayal* words, as indicated by the cluster's placement in more negative ranges. Additionally, *Sanctity/Degradation* exhibits the highest positive bias, whereas *Loyalty/Betrayal* displays the most negative bias. *Authority/Subversion* and *Fairness/Cheating* show nearly neutral biases, with medians around zero.

A more thorough analysis of the distribution can be achieved by focusing on the values for a single semantic axis, specifically the *Authority/Subversion* semantic axis. Figure 6.22 presents a boxplot for each document frame, offering a detailed view of the distribution within each frame. It is clear that this approach reveals differences among the frames. For instance, the mean of the *Cultural Identity* document frame is higher than others, and its overall distribution is more positively skewed compared to other frames. This indicates that articles tagged with the *Cultural Identity* frame tend to use language that leans towards the *virtue* pole of the *Authority/Subversion* semantic axis, effectively using more *authoritarian* words relative to other articles. This phenomenon can be further investigated by analyzing the frames' co-occurrence matrix within the SemEval dataset (see Figure 6.23). The *Cultural Identity* frame occurs 31 times in total, and it co-occurs with the *Morality* frame 21 of those times, more frequently than with any other frame. This frequent co-occurrence highlights a connection between these two frames, which is also evident in their

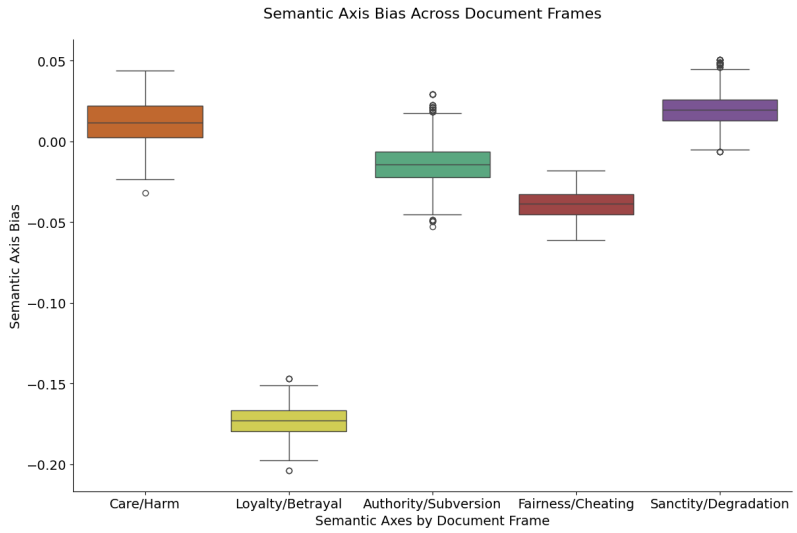

Figure 6.21.: *The chart displays the distribution of semantic axis bias within document frames from the SemEval dataset. The boxplots represent bias values for the five semantic axes: Care/Harm, Loyalty/Betrayal, Authority/Subversion, Fairness/Cheating, and Sanctity/Degradation. The y-axis measures semantic axis bias values, which range from -0.15 to 0.10, with positive values signifying alignment with the first term in each semantic axis pair. This figure highlights the different semantic orientations across various moral foundations within the analyzed documents.*

similar high semantic axis biases. Conversely, articles tagged with the *Security and Defense* frame show a lower mean than all other articles, indicating a predominant use of *vice* words in *Security and Defense* framed articles compared to those not labeled as such. Referring back to Figure 6.22, this pattern is consistent across the other four semantic axes (see Figure B.6 in Appendix B).

Furthermore, the Semantic Axis analysis extends to comprehend the connection between semantic axis biases and the type of articles. As mentioned, an article can be categorized as either an *opinion*, *reporting*, or *satire* piece. Figure 6.24 illustrates a clear pattern where *satire* articles display more positive biases across all semantic axes compared to *opinion* or *reporting* articles. In contrast, the distinction between the *opinion* and *reporting* articles is less evident, as the *reporting* articles exhibit lower biases than the *opinion* articles in the *Authority/Subversion*, *Care/Harm*, and *Loyalty/Betrayal* frames, but not in the *Fairness/Cheating* and *Sanctity/Degradation* frames. This suggests a unique linguistic style in satirical content, which consistently shows higher positive bias across all semantic axes. The mean bias scores for each category provide additional insight into the different semantic orientations of opinion, reporting, and satire articles.

Incorporating semantic axis intensity offers additional insights into the utilization of a semantic axis within a specific target corpus. Figure 6.25 illustrates the correlation between the average bias of each document frame and its average semantic axis intensity for the *Care/Harm* semantic axis. This graph highlights the more negative bias of *Health and Safety* and the more positive semantic axis bias of *Cultural Identity* along the x-axis, while also showing the level of semantic axis intensity on the y-axis. Notably, *Health and Safety* exhibits the highest semantic axis intensity,

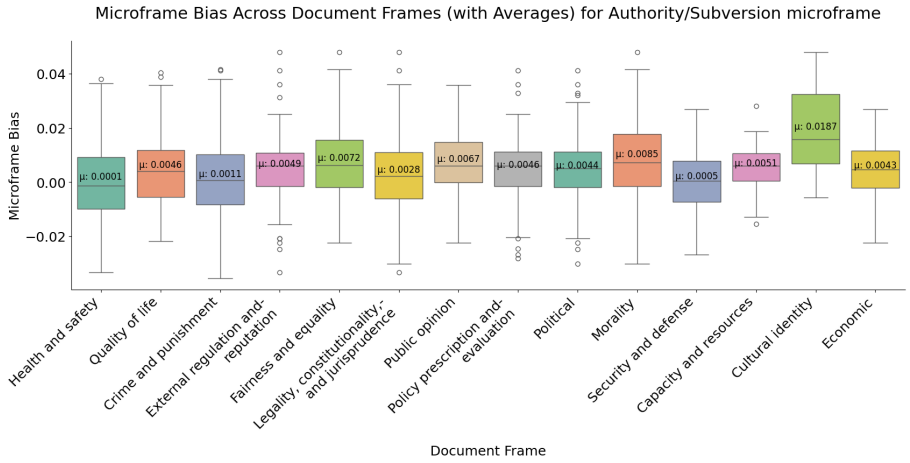

Figure 6.22.: *Authority/Subversion semantic axis bias distribution across document frames in the SemEval dataset. The boxplots present bias values for 15 unique document frames, with the y-axis indicating semantic axis bias values from -0.04 to 0.05. Each distribution's mean bias values (μ) are displayed.* Cultural Identity *shows the highest positive bias ($\mu = 0.0187$), while most frames have slight positive biases.* Health and Safety *exhibits the most neutral bias ($\mu = 0.0001$). This chart highlights subtle differences in the use of authority-related language across various thematic frames in news stories.*

indicating a significant use of the *Care/Harm* semantic axis in these articles. Conversely, articles categorized as *Capacity and Resources* show a much lower intensity, suggesting minimal use of this semantic axis. This pattern is consistent across other semantic axes as well (see Figure B.7 in Appendix B).

By examining not only the document frame but also the article type (see Figure 6.26), it is observed that the overall density distribution of all articles of the *satire* type (shown in yellow) tends to skew towards the right side of the x-axis, indicating a generally positive semantic axis bias. In contrast, *opinion* (depicted in green) and *reporting* (depicted in red) articles overlap significantly, showing minimal differences in leaning. This trend is consistent across all five semantic axes. This plot also highlights the limitations of FrameAxis, as it is challenging to differentiate between more than two classes, and the language used in *opinion* and *reporting* articles is too similar to effectively distinguish between these types. Only *satire* articles are more clearly separated from the rest. The visualizations for the remaining four semantic axes are presented in Figure B.8 located in B.

Analyzing the variations in word-level bias (see Eq: 5.22) and intensity (see Eq: 5.23) demonstrates how specific terms affect framing effects. As depicted in Figure 6.27, the terms that show the most significant bias shifts in articles using the *Economic* frames highlight how particular words can direct the narrative toward certain biases (see Figure B.9 in Appendix B, view online).

Figure 6.23.: *Co-occurrence matrix of frames in the SemEval dataset. The heatmap shows the frequency of frame pairs in annotated articles, with color intensity and numerical values indicating co-occurrence counts. Diagonal elements represent the total occurrences of each frame (e.g., 'Crime and Punishment' appears 227 times). Notable high co-occurrences include 'Crime and Punishment' with 'Morality' (109 instances) and 'Legality, Constitutionality, and Jurisprudence' (130 instances). The 'Political' frame shows strong associations across multiple categories, particularly with itself (235 occurrences). This visualization reveals thematic relationships and common frame combinations in news articles, highlighting the multi-dimensional nature of framing in media discourse.*

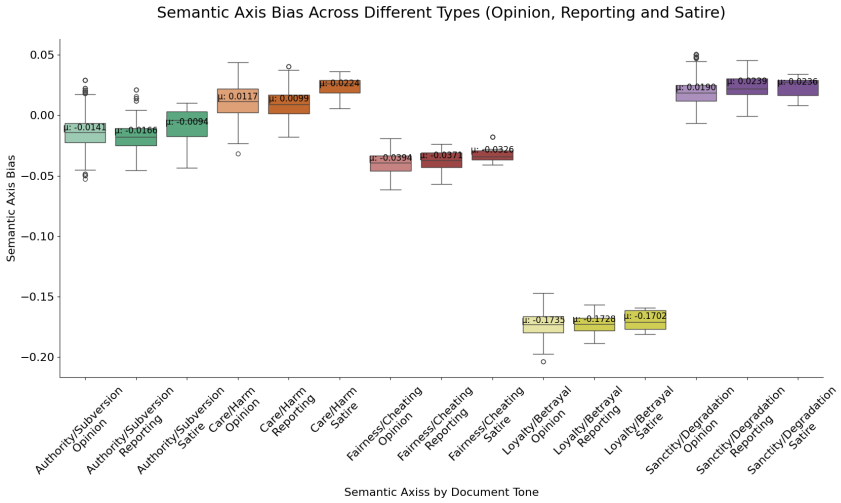

Figure 6.24.: *The dispersion of semantic axis bias among different document categories (opinion, reporting, and satire) in the Media Frames Corpus. The boxplots depict bias scores for five semantic axes (Authority/Subversion, Care/Harm, Fairness/Cheating, Loyalty/Betrayal, and Sanctity/Degradation) within each document classification. The y-axis denotes semantic axis bias scores, where positive values show alignment with the first term in each semantic axis pair. Mean bias scores (μ) are displayed for each distribution.*

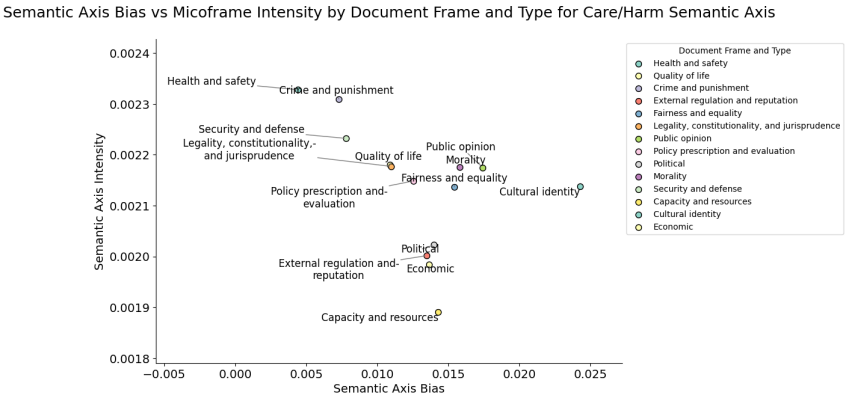

Figure 6.25.: *Correlation between semantic axis bias and intensity for the Care/Harm semantic axis across various document frames in the Media Frames Corpus. The scatter plot shows the average semantic axis bias (x-axis) and intensity (y-axis) for each document frame. Positive bias values indicate a tendency towards the 'Care' aspect, while higher intensity values denote a more frequent application of the Care/Harm semantic axis. Significant differences are noticeable across frames; for instance,* Health and Safety *demonstrates high intensity with moderate bias, whereas* Cultural Identity *shows high bias but lower intensity. This chart illustrates the differing levels and directions of Care/Harm semantic axis usage across document frames, offering insights into framing tactics used across a range of subjects.*

In articles tagged with *Economic* and categorized as *opinion*, the background corpus comprises articles that are non-*opinion and Economic*[9].

[9]Creating this background corpus is important as it provides a reference point for detecting notable differences in the usage of words in the examined articles. By contrasting the specific mix of tones and topics in the current article with the entire dataset (excluding articles with the same combination), we can uncover unique patterns and trends.

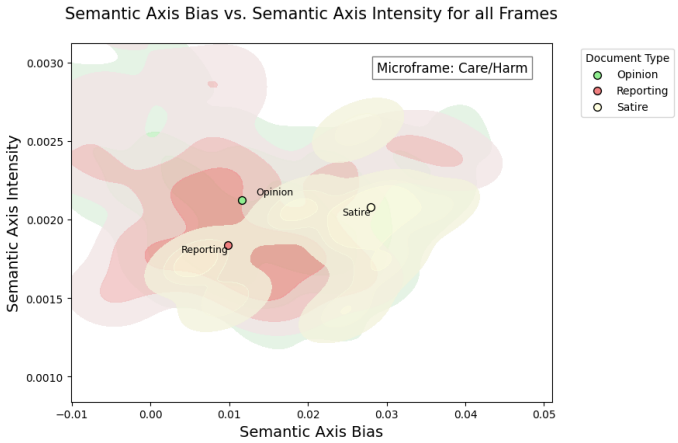

Figure 6.26.: *Visualization of semantic axis bias and intensity for the Care/Harm semantic axis in different document categories (opinion, reporting, and satire) within the Media Frames Corpus. The scatter plot includes density contours to show the concentration of articles per document type. The x-axis denotes semantic axis bias, with positive values aligning with the 'Care' pole, while the y-axis denotes semantic axis intensity. Points represent the average value of the most frequent values for each document category. Reporting articles (red) have lower bias values, opinion pieces (green) shift rightwards with higher bias, and satire (yellow) exhibits the highest bias. This visualization highlights distinct framing patterns in the Care/Harm semantic axis across document categories, indicating systematic framing strategy differences.*

The findings suggest that within the *Fairness/Cheating* semantic axis for *opinion* articles, terms like *obama*, *billion*, *nuclear*, *right*, or *puerto* show a smaller word-level bias shift compared to the background corpus, indicating that these words skew the bias negatively, implying that their presence in *opinion* articles tends to classify the articles more as *cheating* (refer to the center Figure 6.27). Articles categorized as *reporting* also use *obama* and *nuclear* in a more negative context, shifting the bias to the negative pole; the same applies to *sanction*. Additionally, these articles exhibit a higher word-level bias shift for terms like *access* or *bank*, resulting in a shift towards the positive pole (refer to the left Figure 6.27). Conversely, *satire* articles within the same semantic axis use terms like *far*, *conspiracy*, *yellow vests*, *movement*, or *eu*, indicating that these words shift the bias more negatively, towards more *cheating* (refer to the right Figure 6.27).

Similarly, Figure 6.28 explores the extent of these shifts, offering a comprehensive analysis of how terms either enhance or reduce the use of specific frames, thereby influencing the overall narrative intensity (see Figure B.10 in Appendix B, view online). In this scenario, the *Care/Harm* semantic axes in *opinion* articles include words such as *right*, *kavanaugh*[10], and *allegation* within the *Legality, Constitutionality, Jurisdiction* frame, *paddock*[11], *las vegas*[12], and *shooting* within

[10]The term *kavanaugh* refers to Supreme Court Justice Judge *Brett Kavanaugh*.

[11]The term *paddock* refers to *Stephen Craig Paddock*, the mass murderer responsible for killing 58 people during a 2017 music festival attack in Las Vegas [79].

[12]The phrase *las vegas* denotes the city *Las Vegas*, where the mass shooting carried out by Stephen Craig Paddock occurred in 2017 [79].

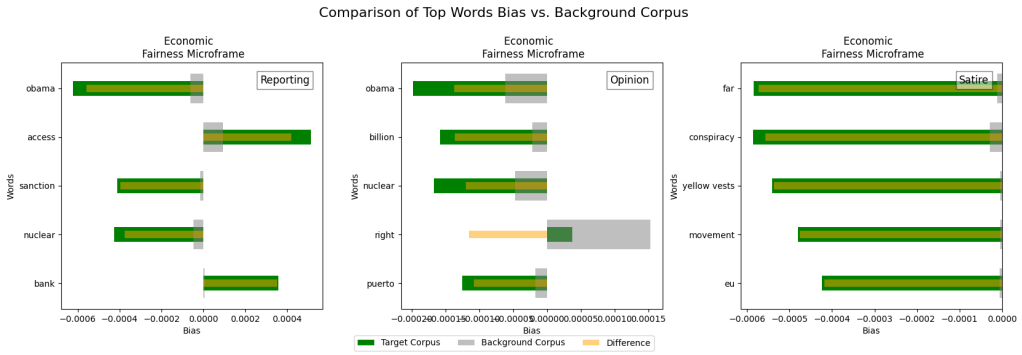

Figure 6.27.: *A comparative examination of word-level bias within the Economic frame of the Fairness/Cheating semantic axis across reporting, opinion, and satire articles in the Media Frames Corpus. The horizontal bar graphs illustrate the bias values of the top five most influential words for each type of document. Green bars signify the bias of the target corpus, gray bars denote the bias of the background corpus, and orange bars show the difference. Positive bias values correspond to the 'Fairness' pole, while negative values correspond to 'Cheating'. This visualization demonstrates distinct patterns of word usage and their associated biases in various types of articles, emphasizing how specific terms shape the framing of economic issues in different journalistic contexts.*

the *Security and Defense* frame, and *good*, *bishop*, and *benedict*[13] within the *Morality* frame as the most notable. These words are intrinsically linked to their respective frames and are logical; for instance, terms like *kavanaugh* and *allegation* are closely tied to legal matters, *paddock* or *shooting* are associated with security issues, while *bishop* and *benedict* are strongly connected to moral topics.

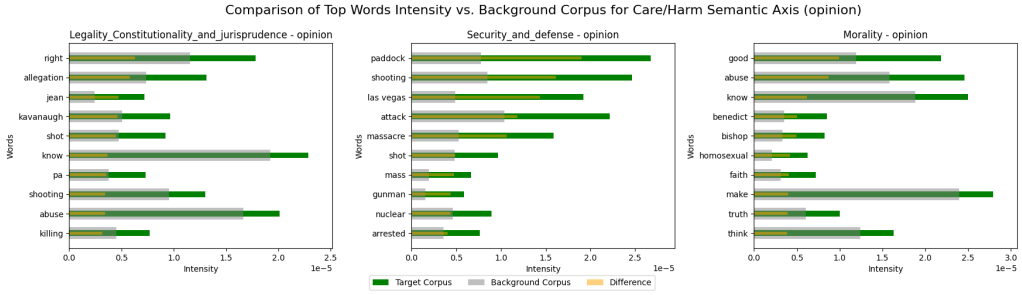

Figure 6.28.: *A comparative evaluation of word-level intensity for the Care/Harm semantic axis in opinion pieces spanning three frames: Legality, Constitutionality, and Jurisprudence; Security and Defense; and Morality. The horizontal bar graphs illustrate the intensity metrics for the top ten most impactful words within each frame. Green bars signify the target corpus intensity, gray bars denote the background corpus intensity, and orange bars represent the difference. Higher intensity measurements indicate a stronger focus on the Care/Harm dimension within the respective frames. This visualization uncovers unique patterns of word employment and their associated intensities, demonstrating how specific terms contribute to framing care and harm issues across various thematic contexts in opinion articles.*

[13]The term *benedict* refers to *Pope Emeritus Benedict XVI.*.

6.3.3. Fine-tuning of MuSE-DLF

The upcoming sections outline the reasoning behind the constraints imposed on dataset preprocessing and provide a detailed explanation of the fine-tuning procedure for the MuSE-DLF model.

Dataset Settings

In Experiment 2, dataset constraints, such as the number of sentences or the length of each sentence, were defined to align with the unique characteristics of the SemEval dataset, while also considering computational limitations. Refer to Table 6.5 for the dataset settings.

Metric	Value
Number of sentences per document	32
Maximum sentence length	64 tokens
Maximum semantic roles per sentence	13
Maximum semantic role length	18 tokens

Table 6.5.: *SemEval Dataset Settings*

The constraints of these datasets, now termed dataset settings, were derived from a thorough analysis of the SemEval dataset, as described in Appendix A in the Tables A.10 and A.11.

The SemEval dataset shows much greater variation in document length compared to the Media Frames Corpus utilized in Experiment 1. As indicated in Table A.10, the 99th percentile for the number of sentences per document is 166, significantly exceeding the previous dataset. Due to memory limitations, we empirically established that 32 sentences per document, which is a multiple of 8, was the maximum feasible while still encompassing more than half of the corpus's articles.

We determined the maximum sentence length to be 64 tokens, which matches our findings that 95% of sentences in the dataset are 56 words or less. To translate words into tokens, we used the word-to-token ratio of 1.35 derived from the SemEval dataset (slightly above the 1.31 ratio in the Media Frames Corpus). Calculating 56 words by 1.35 tokens per word gives roughly 75.6 tokens. Although rounding to the nearest multiple of 8 would result in 80 tokens, memory limitations led us to select the nearest lower multiple of 8, setting the maximum at 64 tokens per sentence[14].

For semantic roles, we maintain 13 roles per sentence, encapsulating 95% of all documents in the corpus (see Table A.11). The upper limit for semantic role length was established at 18 tokens, calculated from the 95th percentile of words per semantic role (14 words) and modified by the 1.35 word-to-token ratio.

The dataset settings are selected to capture the attributes of the SemEval dataset to the greatest extent feasible while remaining computationally efficient.

[14] We utilized a RoBERTa-base tokenizer, achieving an average of 1.35 tokens per word in our dataset. See Appendix B.1 for further details on the tokenization method.

Fine-tuning of MuSE-DLF

The fine-tuning of the MUSE-DLF model utilized insights from the fine-tuning of its predecessor, the SLMuSE-DLF model. Additionally, we employed a parallel configuration of four GPUs to speed up the training process, enabling efficient experimentation with various parameter settings.

Performance evaluation transitioned to using micro and macro F1 scores as the main metrics. This change aligns with SemEval competition standards, enabling easier comparisons with other models. These metrics provide a more detailed assessment of model performance in tasks involving multi-label classification, particularly with imbalanced datasets [88]. The SemEval dataset consists out of three seperate datasets: *dev*, *train* and *test*. The training dataset was used to train the model, the development dataset was used for evaluation after each epoch, and the testing dataset was used for the final model assessment.

Early stopping techniques were used to prevent overfitting and optimize computational resources. The specific criteria for early stopping were adapted to the multi-label classification task and the new evaluation metrics. This included monitoring micro and macro F1 scores on the development set and stopping training if no improvement was observed over a specified number of epochs for the micro F1 training.

At first, we employed PyTorch's BCEWithLogitsLoss, a commonly applied loss function for multi-label tasks [39]. However, we identified that the dataset imbalance along with the frequent false positives during training were causing issues, prompting us to adopt a different loss function. We implemeted the Scaled Asymmetric Loss function (for a detailed explanation, see Appendix B.3.1). This method leverages the PyTorch implementation from Ben-Baruch et al. [9]. The approach tackles class imbalance by asymmetrically handling positive and negative samples, assigning different penalties to false positives and false negatives. Hyperparameters were configured as $\gamma_{neg} = 3$, $\gamma_{pos} = 2$, and $clip = 0.05$, which improved precision without a considerable reduction in recall. To ensure effective balancing with the unsupervised loss, it was necessary to scale the loss. Consequently, we used the asymmetric loss function with a scaling factor of 12, thus adjusting the loss appropriately.

As a learning rate scheduler, we utilize CosineAnnealingWarmRestarts as our learning rate scheduler, configuring the hyperparameters $\mathbf{T_0}$, $\mathbf{T}_{mult}$, and *eta_min* identically to those used for SLMuSE-DLF fine-tuning. To maintain computational efficiency while ensuring sufficient training, the number of epochs was capped at 10.

Table 6.6 displays the hyperparameter ranges that were evaluated and the ultimate settings for the optimal model.

6.3.4. Results and Evaluation

The upcoming sections outline the metrics obtained from the model, analyze and explain these metrics, and evaluate the explainability of the MuSE-DLF model.

Parameter	Range Explored	Final Value
Learning Parameters		
Epochs	10-20	10
Learning Rate	0.00001 - 0.001	0.00007
Batch Size	8 - 32	8
Dropout Probability	0.1 - 0.4	0.3
Optimizer	adam, adamw	adamw
Adam/AdamW Weight Decay	0.0000001 - 0.01	0.000001
Adam/AdamW use AMS Grad	True, False	True
Model Dimensions		
Hidden Dimension	768 - 2048	768
Number of Layers (Unsupervised)	1 - 3	2
Number of Layers (Unsupervised FrameAxis)	1 - 3	2
Activation and Normalization		
Activation Function (Unsupervised)	relu, gelu	relu
Use Layer Norm (Unsupervised)	True, False	True
Activation Function (Unsupervised FrameAxis)	relu, gelu	relu
Use Layer Norm (Unsupervised FrameAxis)	True, False	True
Pooling and Embedding		
Sentence Pooling	mean, cls	mean
Hidden State for Embedding	last, second-to-last	second-to-last
Matmul Input (Unsupervised)	g, d	g
Matmul Input (Unsupervised FrameAxis)	g, d	g
Gumbel Softmax Log (Unsupervised)	True, False	False
Gumbel Softmax Log (Unsupervised FrameAxis)	True, False	False
Loss and Regularization		
Alpha (α)	0.1 - 0.5	0.5
Lambda Orthogonality (λ)	0.001 - 0.1	0.001
Max Margin for Focal Triplet Loss (M)	8	8
Number of Descriptors (t)	8	8
Number of Negative Samples ($N\_$)	32 - 128	64
Asymmetric Loss Negatives (γ_{neg})	1 - 4	3
Asymmetric Loss Positive (γ_{pos})	1 - 4	2
Asymmetric Loss Clip ($clip$)	0.0 - 0.1	0.05
Asymmetric Loss Scaler ($scaler$)	10 - 100	12
Clip Grad Norm ($grad_clip$	0.5 - 2	1
Temperature and Decay		
Tau Decay (τ)	0.0001 - 0.01	0.0005
Minimum Tau (τ_{min})	0.5	0.5

Table 6.6.: *Overview of the Final Hyperparameter Settings for the MuSE-DLF Model and Evaluated Ranges*

Metrics

The performance of MuSE-DLF was evaluated using micro-F1 and macro-F1 scores, consistent with the metrics adopted in the SemEval task. The metrics were calculated with the official scorer provided by the SemEval organizers, enabling comparisons with other competing models. It is crucial to note that the test dataset was solely used to assess the model's performance against the gold labels using this official scorer, thus maintaining the integrity of the evaluation.

Table 6.7 presents a comparison of MuSE-DLF with baseline and other top-performing approaches in the SemEval challenge.

Model	Micro-F1	Macro-F1
SheffieldVeraAI	0.579	0.539
TeamAmpa	0.567	0.510
MarsEclipse	0.562	0.490
MuSE-DLF	0.553	0.497
Hitachi	0.543	0.472
mCPT	0.535	0.482
QUST	0.513	0.462
OCRI	0.513	0.419
BERTastic	0.512	0.446
UMUTeam	0.508	0.415
ACCEPT	0.507	0.502
Baseline	0.350	0.274

Table 6.7.: *Overview of the top 10 models featured in the SemEval-2023 competition, including our introduced MuSE-DLF model. The task organizers provided the baseline for reference. For detailed details on the various models, refer to Piskorski et al. [78].*

During the SemEval 2023 competition, 22 teams took part and submitted their models for evaluation. Among these, MuSE-DLF achieved a micro-F1 score of 0.553 and a macro-F1 score of 0.497, ranking it among the top four models. These scores highlight the strong performance of the model in the competition. Although MuSE-DLF is a monolingual model, unlike some top-performing multilingual models, it provides an advantage in terms of explainability[15]. A comprehensive analysis of the explainability features of MuSE-DLF will be detailed in Section 6.3.4.

Figure 6.29 illustrates the MuSE-DLF training curve, showcasing the patterns in the micro-F1 and macro-F1 scores over the course of the training epochs. The learning curves indicate a consistent improvement in both micro-F1 and macro-F1 scores throughout the training phase. It can be seen that both training and evaluation metrics exhibit a rising trend, reflecting ongoing learning progress. Note that the displayed micro and macro F1 scores are from the *dev* set and do not represent the final model metrics. The model reached its peak performance around epoch 9, and we utilized this model for the final assessment on the holdout *test* dataset. We switch the model to evaluation mode, predict the frames, and compute the final micro and macro F1 scores using the SemEval scorer.

Figure 6.30 illustrates the loss trajectory recorded during the training of MuSE-DLF. This curve presents insights into the model's optimization process. A steady decrease in the loss signifies a continuous improvement in the model's performance. The plotted loss represents the α balanced loss between the supervised and unsupervised losses. The consistent reduction observed in the loss indicates effective learning by the model.

[15] MuSE-DLF's monolingual framework was driven by the lack of multilingual semantic role labeling tools at the time it was created.

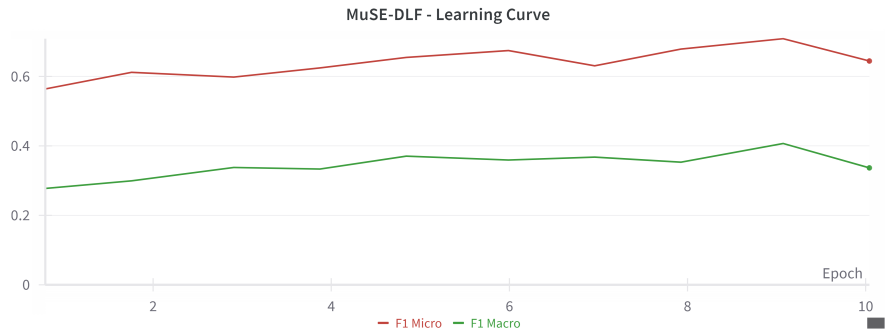

Figure 6.29.: *The MuSE-DLF model's learning curve depicting F1 Micro and F1 Macro scores across 10 epochs. The plot shows a steady enhancement in both metrics, with F1 Micro (red line) consistently surpassing F1 Macro (green line).*

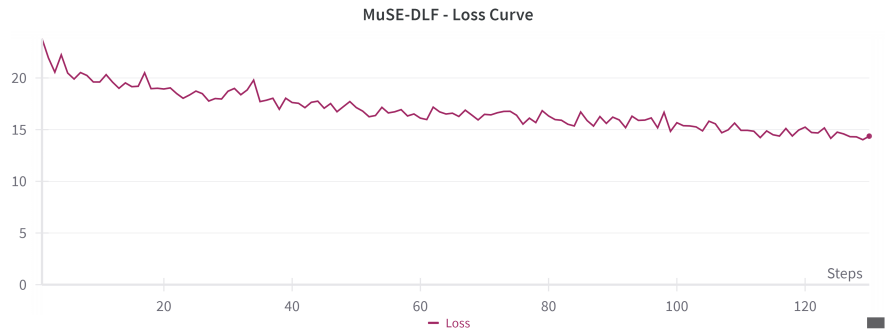

Figure 6.30.: *Training loss curve for the MuSE-DLF model over about 130 steps. The graph exhibits a characteristic decrease in loss, beginning at roughly 24 and falling to around 15 by the conclusion of training. The curve features a steep decline initially, followed by a more moderate reduction, with some variations along the way, indicating continuous learning and the possibility of further improvement with additional training.*

To gain a more detailed understanding of MuSE-DLF performance, we analyzed its F1 scores for each frame category in the *test* datasets. Figures 6.31 and 6.32 illustrate class-wise F1 scores and the distribution of frame occurrences for the *test* data, respectively.

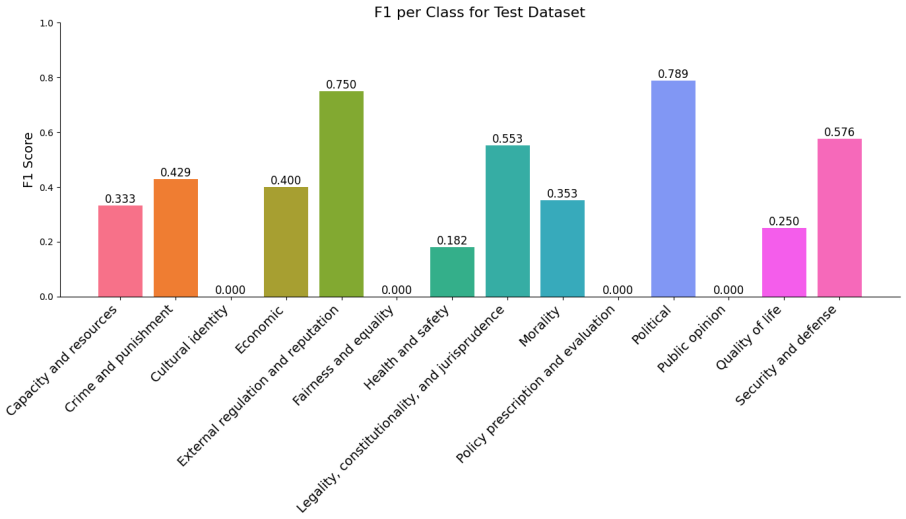

Figure 6.31.: *F1 scores for each frame class in the Test dataset. The chart shows different model performance levels across various frame types. 'Political' frames attain the highest F1 score (0.806), with 'External regulation and reputation' close behind (0.800), demonstrating high precision and recall in these categories. Conversely, some frame types, such as 'Cultural identity', 'Policy prescription and evaluation', and 'Public opinion', have F1 scores of 0.000.*

The evaluation on the *test* set produced a micro-F1 score of 0.553 and a macro-F1 score of 0.497. The minor drop in the micro-F1 score from the development set to the *test* set (0.553) can be explained by a slight decrease in the model's overall performance in common classes when faced with new data. In contrast, the increase in the macro-F1 score (0.406 to 0.497) suggests a more consistent performance across all classes in the *test* set, due to varying class distributions. The reason for the class-specific F1 scores of 0 in *Cultural Identify* and *Public Opinion* is the total lack of articles containing these frames in the *test* dataset.

Explainability

The MuSE-DLF model employs a range of post-hoc explainability techniques that conform to the Explainable AI (XAI) standards detailed in Section 2.2. These techniques seek to offer further understanding of the factors affecting the model's decision-making and the ultimate frame prediction, thereby improving the explainability of the prediction. We suggest three approaches to make the MuSE-DLF model explainable. We use a dataset-level and article-level semantic role analysis based on the work of Khanehzar et al. [53], and we also introduce a novel method to assess the semantic orientation of the articles using data from the semantic axis.

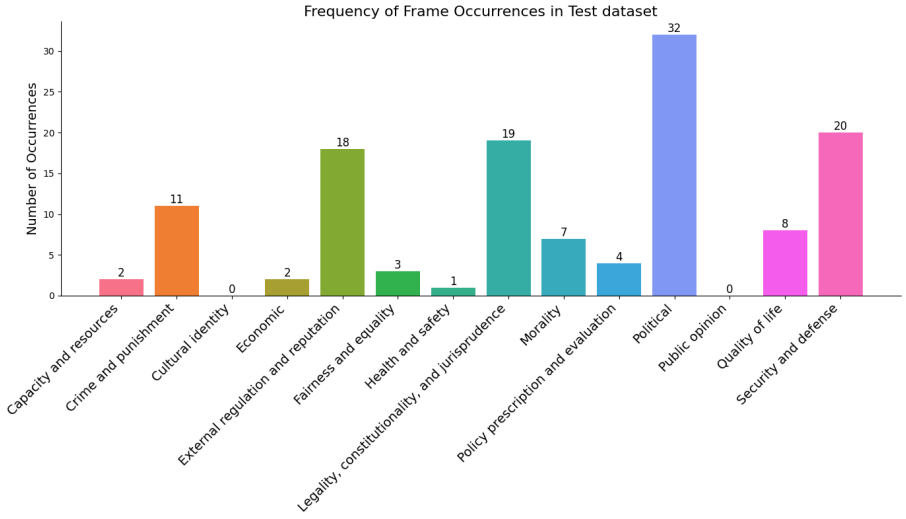

Figure 6.32.: *The figure illustrates the distribution of frame occurrences within the Test dataset. It portrays the frequency of 14 distinct frame categories, highlighting that 'Political' frames are the most frequent (32 occurrences), while categories like 'Cultural identity' and 'Public opinion' have no occurrences (0 occurrences).*

Dataset-level Semantic Role Analysis To gain insights into the most influential words for each frame throughout the entire dataset, we performed a semantic role analysis at the dataset level. We use the values $\mathbf{g}^z$ for each span, where $z \in \{agent, theme, predicate\}$. Following the principle established by Khanehzar et al. [53], we assign a span to a frame if its threshold exceeded 0.8. We then calculate the inverse frequency of each word, remove stop words and other non-meaningful words, and create a table of the most meaningful words for each frame and semantic role.

Table 6.8 provides a selection from this analysis, listing the top 5 words corresponding to each semantic role (predicate, agent, and theme) for four specific frames: *Morality*, *Crime and punishment*, *Economic*, and *Political*. A complete table for all frames can be found in Appendix B Figure B.3.3.

Frame	**Predicate**	**Agent**	**Theme**
Morality	continues, understand, completing, wrote, saying	cardinal, pope, church, francis, priest	god, bishop, pope, field, farrakhan
Crime and punishment	released, arrested, took, found, left	police, attorney, prosecutor, officer, fbi	investigation, report, mueller, evidence, police
Political	continues, agree, avoid, made, confirmed	trump, president, house, government, leader	trump, deal, report, poll, freedom
Economic	deployed, throwing, admitted, signing, denounced	air, austrian, police, delay, american	party, plate, people, ukraine, trump

Table 6.8.: *Top 5 most significant terms identified for the frames Morality, Crime and Punishment, Political, and Economic within the three semantic roles: predicate, agent, and theme.*

This evaluation highlights interesting trends in the language associated with various frames. For example, in the *Morality* frame, religious terms are prevalent for both agents (e.g., *cardinal*,

pope, *church*) and themes (e.g. *god*, *bishop*, *pope*). The predicates for this frame encompass verbs related to communication and comprehension (e.g., *understand*, *wrote*, *saying*).

In contrast, the frame *Crime and Punishment* frame emphasizes legal proceedings and law enforcement. The agent includes terms such as *police, attorney,* and *prosecutor,* while the theme encompasses words such as *investigation, evidence,* and *report.* The verbs in this frame are typically action driven, for instance, *arrested, found,* and *released.*

The *Political* frame is characterized by a distinct emphasis on political figures and procedures. The agent consists of terms such as *trump*, *president*, and *government*, whereas the theme includes words such as *deal*, *poll*, and *freedom*. The predicates in this frame commonly relate to decision-making and communication actions, including *agree*, *avoid*, and *confirmed*.

Interestingly, the *Economic* frame shows less coherent results compared to the other frames. The predicates include a mix of actions (*deployed*, *throwing*, *signing*) and communication verbs (*admitted*, *denounced*). The agent and theme roles also display a diverse range of terms that are not exclusively economic in nature. This discrepancy can be attributed to the relatively low number of articles annotated with the Economic frame in our dataset, as shown in Table A.8 in Appendix B. The *Economic* frame represents only 1.73% of the total articles, which likely results in less robust semantic role identification for this particular frame.

At the dataset level, this semantic role analysis provides valuable understanding of the linguistic patterns associated with different frames. It enables us to identify the typical agents, themes, and actions related to each frame, assisting in decoding the model's predictions and comprehending the complexity of frame classification in news content.

Importantly, although our dataset contains multiple classes, with news articles often incorporating several frames, our method successfully identified intuitive and frame-specific terms for each semantic role-frame link, particularly for frames with higher representation in the dataset. This illustrates the strength of our approach in differentiating between frames, even in complicated, multi-framed contexts. For example, the clear separation of agent terms within the *Morality* frame (e.g., *cardinal*, *pope*) and the *Crime and Punishment* frame (e.g., *police*, *prosecutor*) demonstrates that our model effectively identifies frame-specific semantic roles, even when frames are present simultaneously in the same article.

However, the study reveals deficiencies in our model's performance for less common frames like Economic. This underscores the importance of having balanced datasets for training frame classification models.

Article-level Semantic Role Visualization Building on the explainability of the MUSE-DLF model, this approach visualizes the frame predictions for sentences within particular articles. Unlike dataset-level methods that identify the words most linked to specific frames across an entire corpus, this technique offers clearer insights for single articles. For each sentence, the model analyzes and generates a tensor $\mathbf{d}^z$ for each span, where $z \in \{\textit{predicate, agent, theme}\}$. Every element in this tensor indicates the probability that the span predicted a particular frame. A span

is classified into frame categories when the probability equals or exceeds a threshold. We chose a threshold level of 0.5.

Figure 6.33 illustrates the ability of this visualization method. In this instance, the model predicts the frames *External regulation and reputation*, *Legality Constitutionality and jurisprudence*, *Morality*, *Political*, and *Security and defense*. The true gold labels are *External regulation and reputation*, *Morality*, *Political*, and *Security and defense*. It is observed that four out of five frames were correctly predicted, with only *Legality Constitutionality and jurisprudence* being erroneously included, with a relatively low probability of 0.56.

The visualization of spans in part (a) of Figure 6.33 illustrates that most of the text is colored purple for *Political*, blue for *External regulation and reputation*, red for *Morality*, and brown for *Security and defense*, which matches with the document-level prediction. When dealing with a multi-class model, we do not receive just one prediction per span, but several. Part (b) of the figure offers a more detailed perspective, displaying not only the highest prediction per span but also lower probability predictions. This indicates that some spans, such as *cooperating*, were additionally categorized under *Legality Constitutionality and jurisprudence* (green), likely contributing to its inclusion in the overall prediction. Notably, the spans with multiple annotated frames in (b) consistently match the overall frame predictions, mainly using purple, brown, and blue.

Figure 6.34 illustrates an instance where the advantages of this method become evident. The gold labels for this document are *Crime and punishment*, *Morality*, *Political*, and *Security and defense*. Although the model did not identify *Crime and punishment* as a document-level frame, the span-level visualization shows numerous spans highlighted in blue, which belong to the *Crime and punishment* category. This indicates that span-level analysis can offer valuable insights even when the overall prediction is not entirely accurate, enabling a more detailed understanding of the model's decision-making process.

Figure 6.35 highlights a limitation of the model. In this instance, a substantial number of spans have been tagged as *Political*, resulting in a high-probability prediction for this category at the document level. Nevertheless, this prediction is incorrect when compared to the gold labels of *External regulation and reputation* and *Security and defense*. This instance emphasizes the need for careful evaluation of the model's outputs and the risk of misinterpretation based on span-level categorizations.

The article-level semantic role visualization technique reveals that the span view provides additional understanding into how the model predicts the overall frame, aiding users in comprehending the basis for the model's decisions. Nonetheless, it is important to acknowledge that the model is not flawless, as shown by its performance metrics: a micro-F1 score of 0.553 and a macro-F1 score of 0.497.

Our multi-label assessment shows that the frames designated to specific spans are typically in agreement with the document-level predictions as a whole. This highlights the model's reliability across various analytical levels. This approach offers a detailed perspective on the interaction of

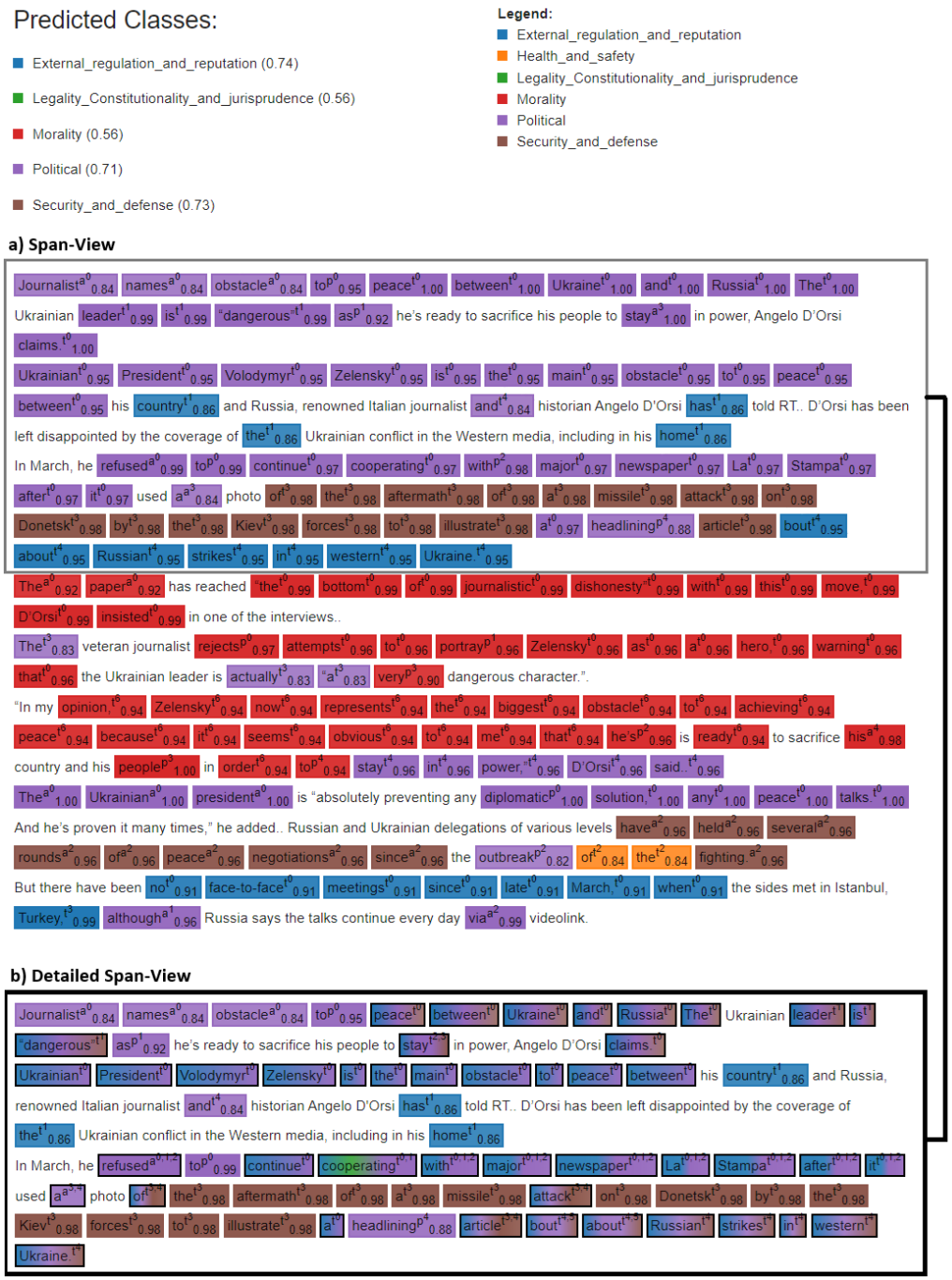

Figure 6.33.: *Visual depictions of semantic roles at the article level. (a) Displays only the primary predicted frame for each semantic role, with frame prediction shown by color. Color intensity and subscript numbers indicate the prediction probability. (b) Illustrates all predicted frames for the opening sentences of the same article by semantic role, using the same color scheme. In both visualizations, prediction probabilities are represented by color intensity and subscript numbers, while the superscript signifies the role related to the semantic role (with p for predicate, a for agent, and t for theme). Connected semantic roles are identified by the superscript number above the semantic role letter.*

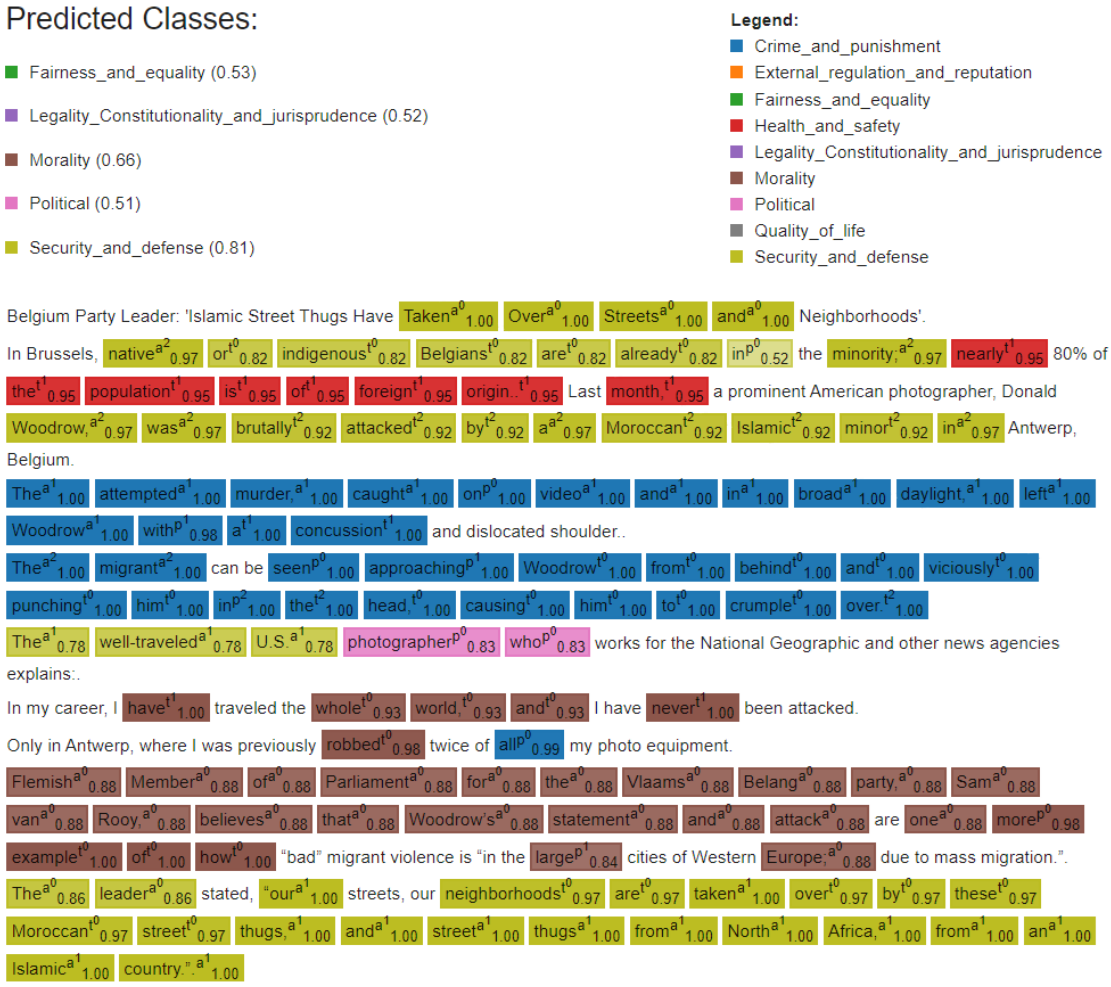

Figure 6.34.: *Illustration of semantic roles for an article where the model's predictions were somewhat correct. The color scheme differentiates various frame categories, with brightness indicating confidence levels in predictions. This case exemplifies how span-level analysis can offer valuable insights even if the predictions at the document level are not fully accurate.*

Figure 6.35.: *An example that illustrates the model's prediction shortcomings is presented. The visualization displays a significant number of spans marked as Political, resulting in a flawed high-probability prediction for this category at the document level.*

multiple frames within a text, providing deeper insights into the complexities of frame prediction and the dynamics of semantic roles.

Semantic Axis Bias Visualization To improve our comprehension of frame predictions and semantic orientations in articles, we utilized a visualization technique grounded in the Semantic Axis values, which is initially presented in this research. This technique provides a detailed perspective on the alignment of each sentence and the entire article with specific semantic axes. We created these visualizations using the values $\mathbf{d}^{fx}$ with a set threshold of 0.5, assigning semantic axis values to a frame when probability exceeded this threshold. These assignments were then used to determine the background bias to which we compare the semantic inclinations of the article.

Figure 6.36 illustrates the assessment of an article using the semantic axis method, focusing on the Care/Harm dimension. Remarkably, the predicted frames for this article (*External regulation and reputation*, *Legality, constitutionality, and jurisprudence*, *Morality*, *Political*, and *Security and defense*) align exactly with the ground truth, highlighting the model's accuracy in multi-class frame classification. This instance exemplifies the model's capability to detect multiple frames within a single article.

The title of the visualization indicates the general semantic inclination of the article. The green bar illustrates the semantic bias of the entire article, whereas the gray bar represents the bias from the background corpus. This background corpus is prepared using predictions from our model.

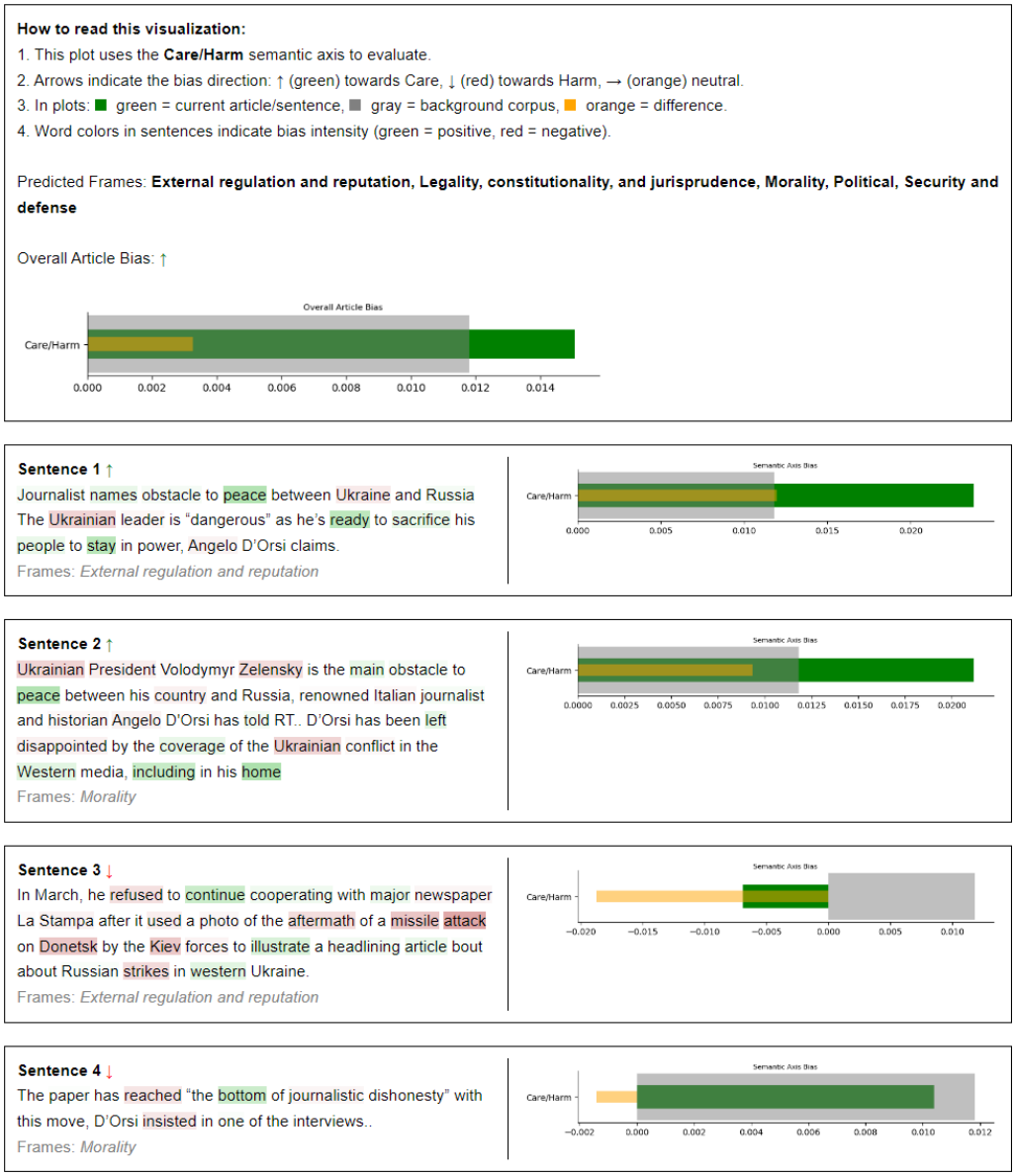

Figure 6.36.: *Illustration of the semantic bias of an article along the Care/Harm axis. The upper part displays the general bias of the article: green for the current article, gray for the corpus background, and orange for the difference. Below that, individual sentence analyses are provided, with word coloring indicating the level of bias (green for positive and red for negative).*

For example, when examining an article marked with an *Economic* frame, the background corpus was created using all articles that were not labeled as *Economic*. This method enables a more detailed comparison between the target article and the broader collection of data.

The contrast between the background (gray) and the current article (green) is shown in orange. A positive difference, marked by the upward-pointing green arrow beside 'Overall Article Bias,' implies that the article favors the positive side of the semantic axis (Care, in this scenario), signifying a higher frequency of caring words as opposed to harming words.

The visual representation extends this analysis to each sentence. Upon examining Sentence 2, we find a positive categorization. The color-coding of each word reveals how they contribute to the analysis. Words such as *peace*, *western*, and *home* are marked as caring (green box), whereas words like *Ukrainian* and *Zelensky* are marked as harming (red box). In this instance, the positive, caring words outweigh the negative ones. Furthermore, we observe the frame predicted by the model for each sentence based on the frame axis values, indicating what the model inferred through the dictionary learning method.

This method allows us to comprehend the framing of articles through their language, even when various frames coexist. For example, the sentence 1 may reflect the author's application of the *External regulation and reputation* frame with an emphasis on caring language, while also incorporating other frames. This multi-frame analytical technique can be applied to study not only the Care/Harm semantic axis but also other semantic axes within each article, offering a thorough perspective of the article's multidimensional framing.

Figure 6.37 extends this analysis to cover all five semantic axes for two different texts. In scenario (a), the model forecasted *Fairness and equality*, *Legality, constitutionality, and jurisprudence*, *Morality*, *Political*, and *Security and defense*, while the actual frames were *Crime and punishment*, *Morality*, *Political*, and *Security and defense*. This leads to two false positives and one false negative. Notably, the prediction at the sentence level using the semantic axis data for *Crime and punishment* aligns with the missing document-level frame, highlighting the model's capability to comprehend nuances that might be overlooked in broader predictions. This illustrates the model's strength in identifying multiple frames on different analysis levels, even if they are not all evident in the final document-level prediction.

The sentence in example (a) shows a negative inclination in every semantic dimension. This might suggest that the author is utilizing a *Crime and punishment* frame, together with terms related to harm, betrayal, subversion, cheating, and degradation.

In Example (b), the model identified the categories *External regulation and reputation*, *Morality*, *Political*, and *Security and defense*, correctly predicting *External regulation and reputation* and *Security and defense*. In this case, the sentence-level predictions are consistent with the document-level predictions, further illustrating the model's ability to maintain coherence across multiple frames and levels of analysis. It should be noted that sentence 3 in this example exhibits a mixed semantic inclination, with a stronger association with betrayal due to the presence of words like *Sweeny* and *great mistake*.

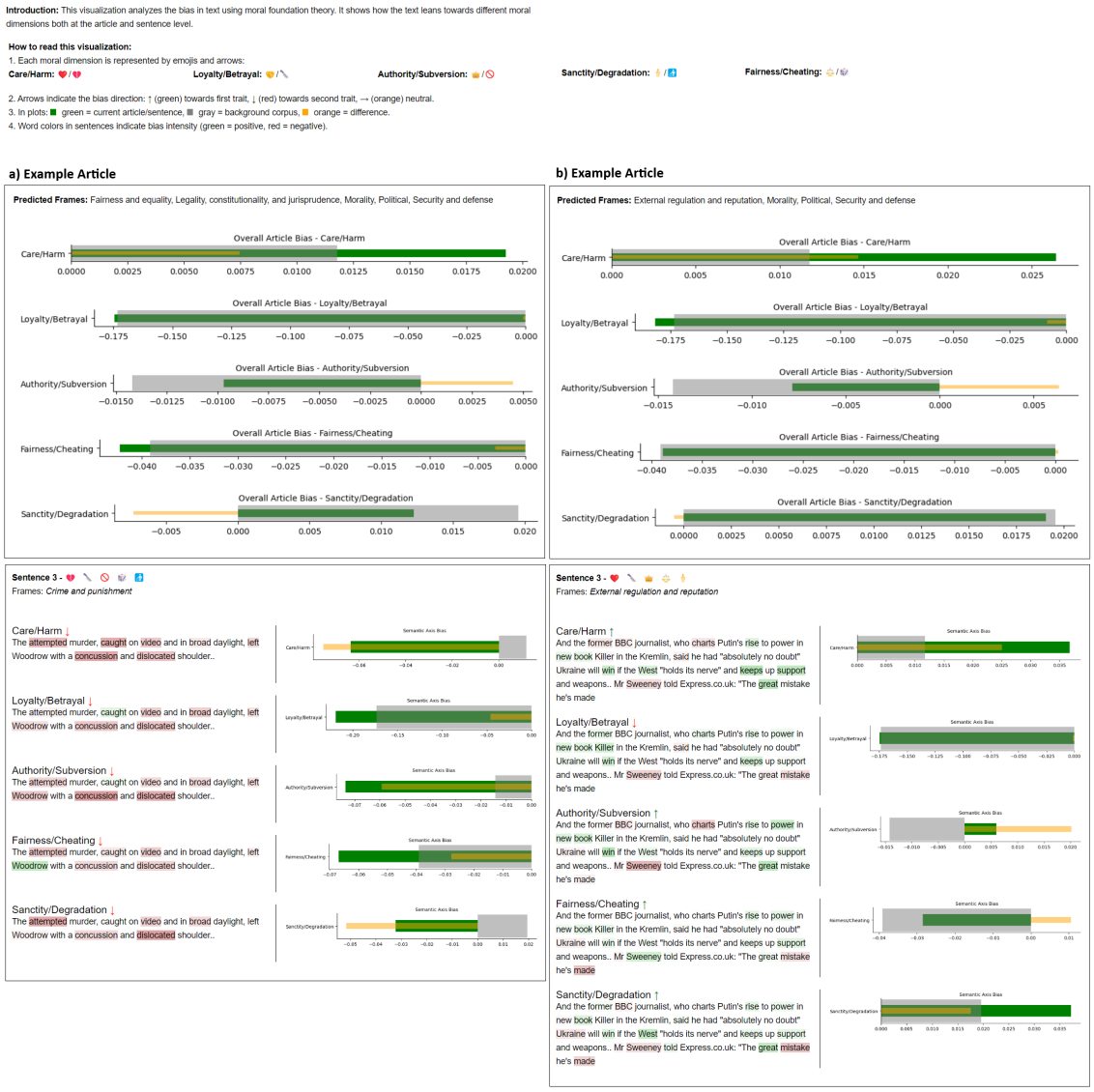

Figure 6.37.: *Evaluation of semantic axis bias in two articles within five moral dimensions. Each subplot illustrates the bias for a specific semantic axis, featuring the overall article bias at the top and sentence-level analysis beneath. The left (a) and right (b) columns depict different articles along with their corresponding predicted frames.*

This visual methodology provides important revelations regarding the connection between word selection and semantic inclinations across defined dimensions in multi-class framing scenarios. It also demonstrates the ability of the model to forecast document-level frameworks based on semantic axis information on a sentence level[16], as shown by the alignment between sentence-level and document-level predictions in both instances.

The depiction of semantic axis bias offers several key benefits in the analysis of multi-class frames. It provides insights at both the article and sentence levels, facilitating a comprehensive understanding of semantic trends across various frames. By integrating data from a background corpus, it places the article's semantic tendencies in the context of the broader dataset, enabling the comparison of multi-frame patterns. The use of color coding for individual terms highlights specific language choices that influence the overall semantic direction, offering a detailed view of how different words affect various frames concurrently.

[16]To measure this alignment, we performed an experiment by comparing sentence-level semantic axis predictions with document-level frames in our test dataset. By using a 0.5 threshold for the semantic axis values, we collected the sentence-level predictions at the document level and compared them with the actual frames of the documents. The analysis showed an overall accuracy of 0.7817, demonstrating a strong correlation between sentence-level semantic axis predictions and document-level frames. The high level of accuracy achieved provides additional evidence of our method's ability to detect multiple document-level frames from sentence-level semantic evaluation

7.

Discussion

Our experiments with the SLMuSE-DLF and MuSE-DLF models produced valuable insights into the quality of our proposed method for frame prediction in news articles. This section discusses these results, focusing on both the performance metrics and the explainability of our models.

7.1. Interpretation of Results

This chapter examines and explains the performance of the proposed models: SLMuSE-DLF and MuSE-DLF.

7.1.1. SLMuSE-DLF Model Performance

The SLMuSE-DLF model, designed for single-label frame classification, demonstrated noteworthy performance on the Media Frames Corpus (MFC). Achieving an accuracy of 0.643 and a macro-F1 score of 0.520. While these metrics did not surpass the FRISS benchmark (accuracy: 0.697, macro-F1: 0.605), they represent a significant advancement over several previous methods, including those by Ji and Smith [47] and Field et al. [26].

The model exhibited varying performance across different frame categories, demonstrated by notably high F1 scores for certain frames like *Political* (0.810) and *Crime and Punishment* (0.750). This indicates that the SLMuSE-DLF model is especially adept at identifying frequently occurring frames in news articles. However, the lower performance on less common frames such as *Fairness and Equality* (F1: 0.250) suggests that the model's effectiveness is somewhat influenced by the prevalence of frames in the training dataset.

Adding the Semantic Axis as a new component seems to improve the model's capability to grasp subtle semantic directions. This is demonstrated by the comprehensive semantic role analysis and bias visualizations along the semantic axis, which shed light on the model's reasoning. These explainability features mark an improvement over earlier non-explainable methods for frame classification.

7.1.2. MuSE-DLF Model Performance

The MuSE-DLF model, designed for multi-label frame classification, demonstrated strong performance on the SemEval-2023 dataset. With a micro-F1 score of 0.553 and a macro-F1 score of 0.497, it secured a place among the top four in the SemEval 2023 competition. This impressive performance is especially significant considering the challenges of multi-label classification and the imbalanced nature of the SemEval dataset.

The model excels in recognizing various frames within a single article, as evidenced by its impressive performance across diverse frame categories. Specifically, the model attained high F1 scores for frames like *Political* (0.806) and *External regulation and reputation* (0.800). This suggests that the MuSE-DLF model proficiently captures the complex nature of framing in news pieces, where multiple frames frequently appear together.

The incorporation of the Semantic Axis component in the multi-label scenario was advantageous, as demonstrated by the model's proficiency in identifying minor variations in framing across different semantic axis dimensions and the significant overlap of predicted frames at the semantic axis level with the document level frames. The semantic axis bias visualizations for multiple frames within a single article demonstrate the model's capacity to provide nuanced insights into the framing strategies employed in complex news narratives.

Nevertheless, the model exhibited limited performance on infrequent frames like *Capacity and Resources* and *Health and Safety* because they were under-represented in the dataset. This underscores the persistent difficulty of creating models capable of generalizing well to rare or unseen frames.

The performance of both the SLMuSE-DLF and MuSE-DLF models showcases the potential of our method to advance automated frame classification. By incorporating semantic role labeling and Semantic Axis components, not only is classification accuracy improved, but the explainability of model predictions is also enhanced.

7.2. Implications for Frame Classification

Our research using the SLMuSE-DLF and MuSE-DLF models reveals important insights for frame classification in Natural Language Processing (NLP). By effectively integrating semantic role labeling with semantic axis components, this method addresses numerous challenges in frame classification, in line with our primary research goals. We have achieved our initial target of developing a new deep learning model that uses semantic role labels and the semantic orientation of texts to improve frame detection performance. The SLMuSE-DLF model, a refined version of the FRISS model employing the FrameAxis technique, has proven its ability to identify subtle framing nuances and effectively predict document level frames. Despite ranking lower than the FRISS model, it outperforms other competitors in single-label frame classification, highlighting the promise of our method.

We have achieved our secondary aim of enhancing frame classification skills to address multi-label situations using the MuSE-DLF model. This progress enables a more accurate representation of the complex nature of framing in real news articles. The MuSE-DLF model's strong performance on the SemEval-2023 dataset, with micro-F1 and macro-F1 scores of 0.553 and 0.497 respectively, demonstrates the model's capability to identify multiple, overlapping frames within an article. This feature is particularly beneficial for investigating intricate issues where various frames typically coexist and interact, marking a significant advance in frame classification.

The introduction and utilization of the SLMuSE-DLF and MuSE-DLF models represent significant advancements in frame classification within NLP. By leveraging semantic data and enabling multi-label classification, they offer a more refined and precise methodology for frame classification. This progress paves the way for further research in NLP by identifying subtle framing elements and multiple frames within a text, thereby allowing for more intricate examinations of content.

7.3. Contributions to Explainable AI in NLP

Our research provides several important additions to the domain of Explainable AI (XAI) in Natural Language Processing (NLP), especially regarding frame classification. These additions specifically meet our third research objective of ensuring post-hoc explainability according to XAI guidelines, and signify a meaningful progress in enhancing the explainability of intricate NLP models.

Integrating semantic role labeling and semantic axis components into our models creates a new foundation for generating explainable predictions in frame classification. In contrast to conventional models that offer limited understanding of their decision-making mechanisms, our approach allows a detailed examination of how particular semantic role words and semantic axes data influence frame predictions. This not only creates model's with high performance but also provides insights into the model's reasoning about framing.

Our method for explainability enhances and broadens existing techniques. The dataset semantic role analysis identifies intuitive terms associated with specific frames throughout the dataset, while the article semantic role analysis delivers detailed insight into sentence-level predictions and their effects on document-level classifications. Both analyses rely on the explainability methods introduced in the FRISS paper. These techniques provide invaluable perspectives on framing structures and the model's decision-making process at various levels of detail.

Our primary advancement in Explainable AI (XAI) within this scope is the deployment of semantic axis analysis at the article level. This novel technique exposes the inclination of articles or sentences towards particular semantic axes, offering insights into the moral foundations and value systems that shape framing strategies. This method provides a richer comprehension of how word usage and their semantic implications affect the overall framing of sentences or articles, significantly enhancing the explainability of our frame classification models.

The visualizations created for sentence-level semantic roles and document-level semantic axis biases present researchers with an intuitive approach to understanding model predictions. Through the use of color-coding for words and phrases according to their semantic roles and biases, we clearly illustrate the influence of word usage on frame classifications.

An essential element of our research is the capability to offer explanations for multi-label classifications. By showcasing the detection and explanation of multiple frames within an article, our work sets the stage for more advanced and refined applications of explainable AI in natural language processing. This is particularly crucial in the domain of news framing, where articles frequently utilize several intersecting frames.

Our research demonstrates a useful approach in the field of XAI for NLP. By developing models that are precise and explainable, we address a significant need in AI. Our method for achieving explainability in frame classification, particularly through the introduction of semantic axis, provides a new direction for making NLP models more understandable. This research could inform the design of explainable AI systems in various natural language processing domains, supporting the goal of creating AI systems that are both reliable and user-friendly.

7.4. Limitations

Although our research progresses the fields of frame classification and explainable AI within NLP, it encounters some limitations. The datasets employed by the SLMuSE-DLF and MuSE-DLF models show an uneven class distribution, with some frames occurring rarely. This disparity negatively impacts model training and performance in less represented classes.

Though the SemEval dataset is smaller than the Media Frames Corpus (MFC), this is not a major issue. However, multi-label classification tasks might gain from having more data. Our models achieve good results with the current data, but expanding the dataset could improve their performance and generalization.

The main reason for the limitations in model fine-tuning was the lack of time. The SLMuSE-DLF model could have been enhanced further by using unlabeled data according to the original FRISS model's approach. Nevertheless, this was skipped due to time constraints.

Due to limitations in computational resources, the MuSE-DLF model had to limit the amount of text processed per document. This included processing fewer sentences and reducing the length of individual sentences. These limitations could potentially impact the accuracy and thoroughness of frame classifications.

Adapting to various languages and fields remains a considerable challenge. The current design of our models specifically targets English-language news articles, limiting their use in tasks involving multiple languages or cross-language frame classification. Their performance in non-news areas, like social media or scholarly papers, has not been thoroughly investigated yet. Moreover, the distinct nature of framing and the potential differences in interpreting semantic axes across various cultural and linguistic backgrounds add further difficulties to their wider application.

A comprehensive ablation study is missing from the Semantic Axis approach, which, despite its novelty and benefits for explainability, introduces some uncertainty about its quantitative impact on model performance in frame detection tasks. Nevertheless, it is important to highlight that enhancing model explainability was the primary objective of incorporating the Semantic Axis approach, rather than merely improving performance measurements.

It is important to note that the MuSE-DLF and SLMuSE-DLF models have not fully realized their potential. Due to the time limitations of this research project, we were unable to perform deeper refinements and optimizations on these models. With additional time and resources, we could perform more comprehensive hyperparameter tuning. This limitation indicates that the current performance of our models, although promising, may only reflect a lower bound of their true capabilities, with considerable room for improvement in future research efforts.

7.5. Future Work

The constraints and results of this research suggest multiple promising avenues for future investigations. Exploring these areas could significantly improve the effectiveness, pertinence, and influence of our frame detection models.

To mitigate imbalances in the dataset, future approaches might investigate methods such as upsampling or downsampling, as well as generating synthetic data. This could lead to better performance of the model on underrepresented classes and improve overall classification accuracy. Also, increasing the dataset size by generating synthetic data might be advantageous, especially for tasks involving multi-label classification. This approach could improve the effectiveness of the models and their ability to generalize.

Further investigations should aim to enhance the models architecture to overcome existing constraints. A crucial avenue for exploration lies in experimenting with diverse semantic axes. By examining a broader spectrum of semantic dimensions, we might discover subtler framing patterns and enhance the model's efficacy as well as its explainability. This could involve devising new semantic axes specific to particular domains or forms of framing, or modifying current psychological or sociological models to establish more thorough semantic representations.

A key area for future research centers on expanding our models to accommodate multiple languages and domains. This requires the development or adjustment of semantic role labelers for languages other than English, which is a significant undertaking that could greatly enhance the global applicability of our approach. Multilingual models designed with explainability could provide important insights into cross-cultural framing patterns.

Additionally, extending the models to various domains outside of news articles, like social media content, scholarly papers, or legal texts, would enhance the versatility of our frame detection method. This would require the development of new annotated datasets for these areas and possibly the adjustment of frame categories to align with the specific features of each domain.

Performing an extensive ablation analysis on the Semantic Axis method would reveal significant insights into its numerical influence on model performance. This may aid in balancing the trade-off between understandability and performance metrics in future versions of the model.

Possible directions for future research emphasize the intricacy of frame classification and point to potential improvements in explainable frame detection techniques. Though there are existing hurdles, our study lays a robust groundwork for ongoing research in this domain.

8.

Conclusion

8.1. Summary of Key Findings

This study has made significant progress in the domain of frame categorization in news media. Our main contribution is the development and analysis of two novel models: SLMuSE-DLF for single-label frame categorization and MuSE-DLF for multi-label frame categorization. These models adeptly combined semantic role labeling and semantic axis elements to improve both the effectiveness and the explainability of frame detection.

The SLMuSE-DLF model was tested on the Media Frames Corpus, achieving an accuracy of 0.643 and a macro-F1 score of 0.520. Though it did not outperform the FRISS baseline, it made notable advancements over many earlier methods. The model excelled in detecting common frames like *Political* and *Crime and Punishment*, but struggled with less common frames.

The MuSE-DLF model demonstrated remarkable performance on the SemEval-2023 dataset, recording a micro-F1 score of 0.553 and a macro-F1 score of 0.497, placing it among the top four models in the competition. This achievement is especially notable due to the intricate nature of multi-label classification and the varied characteristics of the dataset. In addition, MuSE-DLF is the first model capable of predicting multiple frames in news articles, while providing post-hoc explainability.

One of the crucial outcomes of our study was demonstrating the benefits of integrating semantic axis information into frame detection models. This method not only produced accurate classification results, but also made the models' predictions more explainable, shedding light on the semantic leaning of articles.

8.2. Evaluation of Research Goals

Reflecting on the research goals outlined at the beginning of this study, we can evaluate the extent to which each objective has been met:

1. Enhancement of the FRISS Model with Semantic Data: The development of the SLMuSE-DLF model effectively combined semantic role labels and the articles' semantic orientation by utilizing the Semantic Axis approach. Although the model did not exceed the FRISS baseline in overall performance, it showed comparable results and offered improved explainability. The inclusion of data from the semantic axis provided fresh perspectives on the framing process, partially achieving this objective.

2. Development of Multi-Label Frame Classification Capabilities: The objective was thoroughly accomplished through the development of the MuSE-DLF model. The model efficiently managed multi-label classification tasks, evidenced by its excellent results in the SemEval 2023 competition.

3. Ensuring Post-Hoc Explainability: Both the SLMuSE-DLF and MuSE-DLF models included elements designed to improve explainability, which aligns with our XAI objectives. Through semantic role visualization and semantic axis bias analysis, the models offered clear and insightful explanations of their decision-making processes. Although there is still potential for enhancement, this objective was mostly achieved, making a notable contribution to explainable AI in the context of frame classification.

In summary, our research has accomplished substantial advancements towards its main goals. The creation of both single-label and multi-label frame classification models, accompanied by their explainability features, signifies an addition to the field. Nonetheless, the research identified potential areas for further enhancement, especially in improving model performance and extending the methodology to other languages and domains.

8.3. Final Thoughts on the Future of Frame Classification in NLP

The advancement of frame classification in NLP hinges on creating models that are both multilingual and multi-domain. This growth is essential for achieving a comprehensive global insight into media framing in various linguistic and cultural milieus.

Improving the explainability of frame detection models is another important area. Upcoming research should investigate the integration of extra information beyond semantic roles and axes. This might involve emotions or other elements that affect framing, which could enhance both the explainability and accuracy of these models. Explainability in frame detection is essential, as it allows users to comprehend and assess the model's choices, preventing unconscious acceptance of its framing. To improve this further, the discipline should aim at developing models that are easier to interpret.

Although post-hoc explainability methods are beneficial, future research should prioritize the development of deep learning models that incorporate interpretability from the outset. This strategy entails creating models that are fundamentally more transparent in their decision-making pro-

cesses. Methods such as attention visualization [99], or the adoption of interpretable architectures can shed light on how the model considers various aspects of the input when making frame classifications [59].

The optimal future for frame classification in NLP is found at the convergence of high efficiency, explainability, and interpretability. By creating models that excel in performance while offering transparent explanations of their decision-making processes, we can develop tools that significantly enhance our comprehension of framing.

To wrap up, our study has achieved notable advancements in automated frame classification, yet much of the field's potential remains untapped. Looking ahead, it will be crucial to emphasize multi-lingual features, explainability, and interpretability without sacrificing performance. This approach will ensure that frame detection tools serve as effective aids in navigating and interpreting the complex landscape of media framing and other domains, encouraging a more informed public discourse.

Bibliography

[1] Ankur Abhijeet. *Cross-entropy loss - NISER,Bhubaneswar, Odisha*. en. 2023.

[2] Julia Amann et al. "To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems". In: *PLOS Digital Health* 1.2 (Feb. 2022), e0000016. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000016. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9931364/ (visited on 12/23/2023).

[3] Eran Amsalem and Alon Zoizner. "Real, but Limited: A Meta-Analytic Assessment of Framing Effects in the Political Domain". en. In: *British Journal of Political Science* 52.1 (Jan. 2022), pp. 221–237. ISSN: 0007-1234, 1469-2112. DOI: 10.1017/S0007123420000253. URL: https://www.cambridge.org/core/journals/british-journal-of-political-science/article/real-but-limited-a-metaanalytic-assessment-of-framing-effects-in-the-political-domain/4ABFC422A268C55254A03BDE871CF4BB (visited on 06/27/2024).

[4] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. "SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 2450–2461. DOI: 10.18653/v1/P18-1228. URL: https://aclanthology.org/P18-1228/.

[5] Laura Arbelaez Ossa et al. "Re-focusing explainability in medicine". In: *Digital Health* 8 (Feb. 2022), p. 20552076221074488. ISSN: 2055-2076. DOI: 10.1177/20552076221074488. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8841907/ (visited on 12/23/2023).

[6] AWS. *Interpretability versus explainability - Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions*. 2023. URL: https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html (visited on 12/23/2023).

[7] David Bamman and Noah A. Smith. "Open Extraction of Fine-Grained Political Statements". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Lan-*

guage Processing (2015), pp. 76–85. DOI: 10.18653/v1/D15-1008. URL: https://aclanthology.org/D15-1008/.

[8] Gregory Bateson. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. en. Chicago, IL: University of Chicago Press, 1972. ISBN: 978-0-226-03905-3. URL: https://press.uchicago.edu/ucp/books/book/chicago/S/bo3620295.html (visited on 12/23/2023).

[9] Emanuel Ben-Baruch et al. *Asymmetric Loss For Multi-Label Classification*. arXiv:2009.14119 [cs]. July 2021. DOI: 10.48550/arXiv.2009.14119. URL: http://arxiv.org/abs/2009.14119 (visited on 08/09/2024).

[10] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems* 29 (2016).

[11] Amber Boydstun et al. *Tracking the Development of Media Frames within and across Policy Issues*. 2014. URL: https://kilthub.cmu.edu/articles/journal_contribution/Tracking_the_Development_of_Media_Frames_within_and_across_Policy_Issues/6473780.

[12] Amber E. Boydstun and Justin H. Gross. "Identifying Media Frames and Frame Dynamics Within and Across Policy Issues". In: 2013. URL: https://www.semanticscholar.org/paper/Identifying-Media-Frames-and-Frame-Dynamics-Within-Boydstun-Gross/e06095c181afe22f1dba486e21928b48d0f74764 (visited on 03/18/2024).

[13] Dallas Card et al. "Analyzing Framing through the Casts of Characters in the News". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Ed. by Jian Su, Kevin Duh, and Xavier Carreras. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1410–1420. DOI: 10.18653/v1/D16-1148. URL: https://aclanthology.org/D16-1148 (visited on 08/18/2024).

[14] Dallas Card et al. "The Media Frames Corpus: Annotations of Frames Across Issues". In: *undefined* (2015). URL: https://www.semanticscholar.org/paper/The-Media-Frames-Corpus%3A-Annotations-of-Frames-Card-Boydstun/92408cc19033cc4af29accef3793014ab79355c2.

[15] Michael J. Carter. "The Hermeneutics of Frames and Framing: An Examination of the Media's Construction of Reality". en. In: *Sage Open* 3.2 (Apr. 2013). Publisher: SAGE Publications, p. 2158244013487915. ISSN: 2158-2440. DOI: 10.1177/2158244013487915. URL: https://doi.org/10.1177/2158244013487915 (visited on 06/27/2024).

[16] Brendan Chambers and James Evans. "Semantic maps and metrics for science Semantic maps and metrics for science using deep transformer encoders". In: *undefined* (2021). URL: https://www.semanticscholar.org/paper/Semantic-maps-and-metrics-for-science-Semantic-maps-Chambers-Evans/8ecb08beba33df69eaa7a13f4dbed0390e47415e.

[17] Fei-Fei Cheng and Chin-Shan Wu. "Debiasing the framing effect: The effect of warning and involvement". In: *Decision Support Systems* 49.3 (June 2010), pp. 328–334. ISSN: 0167-9236. DOI: 10.1016/j.dss.2010.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0167923610000709 (visited on 06/27/2024).

[18] Dennis Chong and James N. Druckman. "A Theory of Framing and Opinion Formation in Competitive Elite Environments". en. In: *Journal of Communication* 57.1 (2007). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.2006.00331.x, pp. 99–118. ISSN: 1460-2466. DOI: 10.1111/j.1460-2466.2006.00331.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2006.00331.x (visited on 11/18/2023).

[19] Frank R. Cicero. "Behavioral Ethics: Ethical Practice Is More Than Memorizing Compliance Codes". In: *Behavior Analysis in Practice* 14.4 (June 2021), pp. 1169–1178. ISSN: 1998-1929. DOI: 10.1007/s40617-021-00585-5. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8586372/ (visited on 12/05/2023).

[20] Susan E. Defranzo. "Advantages and Disadvantages of Surveys". In: *Snap Surveys* (Nov. 2012). URL: https://www.snapsurveys.com/blog/advantages-disadvantages-surveys/.

[21] James N. Druckman. "The limits of political manipulation: Psychological and strategic determinants of framing". Ph.D. Ann Arbor: University of California, San Diego, 1999. URL: https://www.proquest.com/dissertations-theses/limits-political-manipulation-psychological/docview/304497873/se-2?accountid=14570.

[22] University of Edinburgh. *Media framing and how it can shift the narrative*. en. URL: https://www.business-school.ed.ac.uk/about/news/media-framing-and-how-it-shifts-the-narrative (visited on 12/05/2023).

[23] Hatice Ekici, Emine Yücel, and Sevim Cesur. "Deciding between moral priorities and COVID-19 avoiding behaviors: A moral foundations vignette study". In: *Current Psychology (New Brunswick, N.J.)* (2021), pp. 1–17. DOI: 10.1007/s12144-021-01941-y. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8173317/.

[24] Robert M. Entman. "Framing: Toward Clarification of a Fractured Paradigm". en. In: *Journal of Communication* 43.4 (1993). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1993.tb01304.x, pp. 51–58. ISSN: 1460-2466. DOI: 10.1111/j.1460-2466.1993.tb01304.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1993.tb01304.x (visited on 11/18/2023).

[25] Bernhard Ertl, Florian G. Hartmann, and Jörg-Henrik Heine. "Analyzing Large-Scale Studies: Benefits and Challenges". In: *Frontiers in Psychology* 11 (2020). ISSN: 1664-1078. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.577410 (visited on 12/21/2023).

[26] Anjalie Field et al. "Framing and Agenda-setting in Russian News: a Computational Analysis of Intricate Political Strategies". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 3570–3580. DOI: 10.18653/v1/D18-1393. URL: https://aclanthology.org/D18-1393 (visited on 08/18/2024).

[27] Charles J. Fillmore. "The case for case". en. In: *Microprocessors and Microsystems*. Vol. 10. ISSN: 01419331 Issue: 8 Journal Abbreviation: Microprocessors and Microsystems. Oct. 1986, p. 451. DOI: 10.1016/0141-9331(86)90222-X. URL: https://linkinghub.elsevier.com/retrieve/pii/014193318690222X (visited on 12/23/2023).

[28] Kary Främling. *Feature Importance versus Feature Influence and What It Signifies for Explainable AI*. arXiv:2308.03589 [cs]. Aug. 2023. DOI: 10.48550/arXiv.2308.03589. URL: http://arxiv.org/abs/2308.03589 (visited on 12/23/2023).

[29] Dean Fulgoni et al. "An Empirical Exploration of Moral Foundations Theory in Partisan News Sources". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (2016), pp. 3730–3736. URL: https://aclanthology.org/L16-1591/.

[30] Matt Gardner et al. *AllenNLP: A Deep Semantic Natural Language Processing Platform*. Mar. 2018. URL: https://arxiv.org/pdf/1803.07640.

[31] Devottam Gaurav and Sanju Tiwari. "Interpretability Vs Explainability: The Black Box of Machine Learning". In: *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. Feb. 2023, pp. 523–528. DOI: 10.1109/ICCoSITE57641.2023.10127717. URL: https://ieeexplore.ieee.org/document/10127717 (visited on 12/23/2023).

[32] Erving Goffman. *Frame analysis: An essay on the organization of experience*. 3. print. Cambridge, Mass.: Harvard Univ. Press, 1976. ISBN: 0-674-31656-8.

[33] Jae Sik Ha and Donghee Shin. "Framing the Arab Spring: Partisanship in the news stories of Korean Newspapers". In: *International Communication Gazette* 78.6 (2016), pp. 536–556. ISSN: 1748-0485. DOI: 10.1177/1748048516640213.

[34] Jonathan Haidt and Jesse Graham. "When Morality Opposes Justice: Conservatives Have Moral Intuitions that Liberals may not Recognize". In: *Social Justice Research* 20.1 (2007), pp. 98–116. ISSN: 1573-6725. DOI: 10.1007/s11211-007-0034-z. URL: https://link.springer.com/article/10.1007/s11211-007-0034-z.

[35] Maram Hasanain et al. "QCRI at SemEval-2023 Task 3: News Genre, Framing and Persuasion Techniques Detection Using Multilingual Models". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1237–1244. DOI: 10.18653/v1/2023.semeval-1.172. URL: https://aclanthology.org/2023.semeval-1.172 (visited on 11/23/2023).

[36] Adeel Hassan. *AdeelH/pytorch-multi-class-focal-loss*. original-date: 2020-09-03T09:08:36Z. July 2024. URL: https://github.com/AdeelH/pytorch-multi-class-focal-loss (visited on 07/30/2024).

[37] Vincent F. Hendricks and Mads Vestergaard. "Alternative Facts, Misinformation, and Fake News". en. In: *Reality Lost: Markets of Attention, Misinformation and Manipulation*. Ed. by Vincent F. Hendricks and Mads Vestergaard. Cham: Springer International Publishing, 2019, pp. 49–77. ISBN: 978-3-030-00813-0. DOI: 10.1007/978-3-030-00813-0_4. URL: https://doi.org/10.1007/978-3-030-00813-0_4 (visited on 12/05/2023).

[38] Frederic R. Hopp et al. "The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text". en. In: *Behavior Research Methods* 53.1 (Feb. 2021), pp. 232–246. ISSN: 1554-3528. DOI: 10.3758/s13428-020-01433-0. URL: https://doi.org/10.3758/s13428-020-01433-0 (visited on 04/25/2024).

[39] De Jun Huang. *Learning Day 57/Practical 5: Loss function — CrossEntropyLoss vs BCELoss in Pytorch; Softmax vs...* en. June 2021. URL: https://medium.com/dejunhuang/learning-day-57-practical-5-loss-function-crossentropyloss-vs-bceloss-in-pytorch-softmax-vs-bd866c8a0d23 (visited on 08/12/2024).

[40] Austin Hubner and Graham Dixon. "Is news media sharing an active framing process? Examining whether individual tweets retain news media frames about climate change". In: *Human Communication Research* 49.1 (Jan. 2023), pp. 75–84. ISSN: 1468-2958. DOI: 10.1093/hcr/hqac025. URL: https://doi.org/10.1093/hcr/hqac025 (visited on 12/05/2023).

[41] Huggingface. *FacebookAI/roberta-base · Hugging Face*. Mar. 2024. URL: https://huggingface.co/FacebookAI/roberta-base (visited on 04/30/2024).

[42] Huggingface. *Perplexity of fixed-length models*. English. Apr. 2024. URL: https://huggingface.co/docs/transformers/perplexity (visited on 04/25/2024).

[43] Huggingface. *Summary of the tokenizers*. June 2024. URL: https://huggingface.co/docs/transformers/en/tokenizer_summary (visited on 08/08/2024).

[44] Chip Huyen. *Evaluation Metrics for Language Modeling*. en. Oct. 2019. URL: https://thegradient.pub/understanding-evaluation-metrics-for-language-models/ (visited on 04/25/2024).

[45] Shanto Iyengar. *Is anyone responsible? How television frames political issues*. American politics and political economy series. Chicago: University of Chicago Press, 1991. ISBN: 978-0-226-38855-7. DOI: 10.7208/chicago/9780226388533.001.0001.

[46] Shanto Iyengar and Adam Simon. "News Coverage of the Gulf Crisis and Public Opinion". In: *Communication Research* 20.3 (1993), pp. 365–383. ISSN: 0093-6502. DOI: 10.1177/009365093020003002.

[47] Yangfeng Ji and Noah A. Smith. "Neural Discourse Structure for Text Categorization". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 996–1005. DOI: 10.18653/v1/P17-1092. URL: https://aclanthology.org/P17-1092 (visited on 08/18/2024).

[48] Kristen Johnson and Dan Goldwasser. "Classification of Moral Foundations in Microblog Political Discourse". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2018), pp. 720–730. DOI: 10.18653/v1/P18-1067. URL: https://aclanthology.org/P18-1067/.

[49] Daniel Kahneman and Amos Tversky. "Choices, values, and frames". In: *American Psychologist* 39.4 (1984), pp. 341–350. ISSN: 0003-066X. DOI: 10.1037/0003-066X.39.4.341.

[50] Shigeki Karita et al. "A Comparative Study on Transformer vs RNN in Speech Applications". In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. arXiv:1909.06317 [cs, eess]. Dec. 2019, pp. 449–456. DOI: 10.1109/ASRU46091.2019.9003750. URL: http://arxiv.org/abs/1909.06317 (visited on 04/27/2024).

[51] Sandra Kaufman, Michael Elliott, and Deborah Shmueli. *Frames, Framing and Reframing*. en. Text. Last Modified: 2024-01-12T10:26-07:00. 2003. URL: https://www.beyondintractability.org/essay/framing (visited on 06/27/2024).

[52] Shima Khanehzar, Andrew Turpin, and Gosia Mikolajczak. "Modeling Political Framing Across Policy Issues and Contexts". In: *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*. Ed. by Meladel Mistica, Massimo Piccardi, and Andrew MacKinlay. Sydney, Australia: Australasian Language Technology Association, 2019, pp. 61–66. URL: https://aclanthology.org/U19-1009 (visited on 08/18/2024).

[53] Shima Khanehzar et al. *Framing Unpacked: A Semi-Supervised Interpretable Multi-View Model of Media Frames*. Apr. 2021. URL: https://arxiv.org/pdf/2104.11030.

[54] Guido Kiell. "Entscheidungs-Frames und Framing-Effekte: Varianten, Wirkungen und psychologische Ursachen". de. In: (2019). URL: https://kups.ub.uni-koeln.de/9738/1/Guido_Kiell_Entscheidungs-Frames_und_Framing-Effekte_20190703.pdf.

[55] Donald R. Kinder and Lynn M. Sanders. *Divided by Color: Racial Politics and Democratic Ideals*. en. American Politics and Political Economy Series. Chicago, IL: University of Chicago Press, Oct. 1997. ISBN: 978-0-226-43574-9. URL: https://press.uchicago.edu/ucp/books/book/chicago/D/bo3620441.html (visited on 12/23/2023).

[56] Haewoon Kwak, Jisun, and Yong-Yeol Ahn. *A Systematic Media Frame Analysis of 1.5 Million New York Times Articles from 2000 to 2017*. 2020. URL: https://arxiv.org/pdf/2005.01803.

[57] Haewoon Kwak et al. "FrameAxis: characterizing microframe bias and intensity with word embedding". In: *PeerJ. Computer science* 7 (2021), e644. DOI: 10.7717/peerj-cs.644. URL: https://arxiv.org/pdf/2002.08608.

[58] George Lawton. *UX defines chasm between explainable vs. interpretable AI | TechTarget*. en. URL: https://www.techtarget.com/searchenterpriseai/feature/UX-defines-chasm-between-explainable-vs-interpretable-AI (visited on 12/23/2023).

[59] Xuhong Li et al. *Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond*. en. arXiv:2103.10689 [cs]. July 2022. URL: http://arxiv.org/abs/2103.10689 (visited on 08/18/2024).

[60] Qisheng Liao, Meiting Lai, and Preslav Nakov. "MarsEclipse at SemEval-2023 Task 3: Multi-lingual and Multi-label Framing Detection with Contrastive Learning". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 83–87. DOI: 10.18653/v1/2023.semeval-1.10. URL: https://aclanthology.org/2023.semeval-1.10 (visited on 11/21/2023).

[61] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. arXiv:1708.02002 [cs]. Feb. 2018. DOI: 10.48550/arXiv.1708.02002. URL: http://arxiv.org/abs/1708.02002 (visited on 07/30/2024).

[62] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *CoRR* abs/1907.11692 (2019). _eprint: 1907.11692. URL: http://arxiv.org/abs/1907.11692.

[63] Teemu Maatta. *Natural Language Processing (NLP)*. en. Feb. 2022. URL: https://tmmtt.medium.com/natural-language-processing-nlp-dc2c1d8d4110 (visited on 04/25/2024).

[64] Sadhika Malladi et al. *A Kernel-Based View of Language Model Fine-Tuning*. arXiv:2210.05643 [cs]. June 2023. DOI: 10.48550/arXiv.2210.05643. URL: http://arxiv.org/abs/2210.05643 (visited on 04/25/2024).

[65] Media Matters. *How Fox frames an immigration story for two different audiences: http://t.co/JuoAtEDBFX*. en. Tweet. Aug. 2014. URL: https://twitter.com/mmfa/status/497856477802278912 (visited on 12/05/2023).

[66] Michael Merry, Pat Riddle, and Jim Warren. "A mental models approach for defining explainable artificial intelligence". In: *BMC Medical Informatics and Decision Making* 21.1 (Dec. 2021), p. 344. ISSN: 1472-6947. DOI: 10.1186/s12911-021-01703-7. URL: https://doi.org/10.1186/s12911-021-01703-7 (visited on 12/23/2023).

[67] Negar Mokhberian et al. *Moral Framing and Ideological Bias of News*. 2020. DOI: 10.1007/978-3-030-60975-7âĂŰ16. URL: https://arxiv.org/pdf/2009.12979.

[68] Pedro Ramaciotti Morales and Gabriel Muñoz Zolotoochin. *Measuring the accuracy of social network ideological embeddings using language models*. 2022. URL: https://halshs.archives-ouvertes.fr/hal-03385061/.

[69] Kashif Munir, Hai Zhao, and Zuchao Li. *Neural Unsupervised Semantic Role Labeling*. arXiv:2104.09047 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2104.09047. URL: http://arxiv.org/abs/2104.09047 (visited on 12/26/2023).

[70] Nona Naderi and Graeme Hirst. "Classifying Frames at the Sentence Level in News Articles". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Ed. by Ruslan Mitkov and Galia Angelova. Varna, Bulgaria: INCOMA Ltd., Sept. 2017, pp. 536–542. DOI: 10.26615/978-954-452-049-6_070. URL: https://doi.org/10.26615/978-954-452-049-6_070 (visited on 11/23/2023).

[71] Thomas E. Nelson, Rosalee A. Clawson, and Zoe M. Oxley. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance". In: *American Political Science Review* 91.3 (1997), pp. 567–583. ISSN: 0003-0554. DOI: 10.2307/2952075.

[72] Milda Norkute et al. "Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization". In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI EA '21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–7. ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3443441. URL: https://doi.org/10.1145/3411763.3443441 (visited on 12/23/2023).

[73] Kristen Olson, James Wagner, and Raeda Anderson. "Survey Costs: Where are We and What is the Way Forward?" In: *Journal of Survey Statistics and Methodology* 9.5 (Nov. 2021), pp. 921–942. ISSN: 2325-0984. DOI: 10.1093/jssam/smaa014. URL: https://doi.org/10.1093/jssam/smaa014 (visited on 12/21/2023).

[74] Michael Oswald. *Strategisches Framing: Eine Einführung*. de. Wiesbaden: Springer Fachmedien, 2019. ISBN: 978-3-658-24283-1 978-3-658-24284-8. DOI: 10.1007/978-3-658-24284-8. URL: http://link.springer.com/10.1007/978-3-658-24284-8 (visited on 12/23/2023).

[75] Cecilia Panigutti et al. "The role of explainable AI in the context of the AI Act". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. New York, NY, USA: Association for Computing Machinery, June 2023, pp. 1139–1150. ISBN: 9798400701924. DOI: 10.1145/3593013.3594069. URL: https://dl.acm.org/doi/10.1145/3593013.3594069 (visited on 12/23/2023).

[76] Souneil Park et al. "The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns". In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. Ed. by Pamela Hinds. ACM Conferences. New York, NY: ACM, 2011, p. 113. ISBN: 978-1-4503-0556-3. DOI: 10.1145/1958824.1958842.

[77] Jiaxi Peng et al. "Five Different Types of Framing Effects in Medical Situation: A Preliminary Exploration". In: *Iranian Red Crescent Medical Journal* 15.2 (Feb. 2013), pp. 161–165. ISSN: 2074-1804. DOI: 10.5812/ircmj.8469. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3652505/ (visited on 06/27/2024).

[78] Jakub Piskorski et al. "SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup". en. In: *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 2343–2361. DOI: 10.18653/v1/2023.semeval-1.317. URL: https://aclanthology.org/2023.semeval-1.317 (visited on 11/21/2023).

[79] Associated Press. *Vegas gunman inspired by criminal father's reputation*. en. Jan. 2019. URL: https://www.nbcnews.com/storyline/las-vegas-shooting/vegas-gunman-stephen-paddock-inspired-criminal-father-s-reputation-n964066 (visited on 06/17/2024).

[80] J. Priniski et al. "Mapping Moral Valence of Tweets Following the Killing of George Floyd". In: *undefined* (2021). URL: https://www.semanticscholar.org/paper/Mapping-Moral-Valence-of-Tweets-Following-the-of-Priniski-Mokhberian/690a9413993c30acfd36735e4a2a5b017b385e61.

[81] Tommaso Radicioni et al. "Networked partisanship and framing: A socio-semantic network analysis of the Italian debate on migration". In: *PloS one* 16.8 (2021), e0256705. DOI: 10.1371/journal.pone.0256705.

[82] Ashwin Rao et al. "Political Partisanship and Antiscience Attitudes in Online Discussions About COVID-19: Twitter Content Analysis". In: *Journal of medical Internet research* 23.6 (2021), e26692. ISSN: 1438-8871. DOI: 10.2196/26692. URL: https://pubmed.ncbi.nlm.nih.gov/34014831/.

[83] Markus Reiter-Haas, Simone Kopeinik, and Elisabeth Lex. "Studying Moral-based Differences in the Framing of Political Tweets". In: *Proceedings of the International AAAI Conference on Web and Social Media Vol. 15 (* (2021). URL: https://arxiv.org/pdf/2103.11853.

[84] Luke Salomone. *What is Perplexity?* en. Section: posts. Apr. 2021. URL: https://lukesalamone.github.io/posts/perplexity/ (visited on 04/25/2024).

[85] Max Schemmer et al. *On the Influence of Explainable AI on Automation Bias*. arXiv:2204.08859 [cs]. Apr. 2022. DOI: 10.48550/arXiv.2204.08859. URL: http://arxiv.org/abs/2204.08859 (visited on 12/23/2023).

[86] Da Scheufele. "Framing as a theory of media effects". en. In: *Journal of Communication* 49.1 (1999). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1999.tb02784.x, pp. 103–122. ISSN: 1460-2466. DOI: 10.1111/j.1460-2466.1999.tb02784.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1999.tb02784.x (visited on 12/23/2023).

[87] Matheus Schmitz, Goran Murić, and Keith Burghardt. *Detecing Anti-Vaccine Users on Twitter*. 2021. URL: https://arxiv.org/pdf/2110.11333.

[88] scikit-learn. *f1_score*. en. 2024. URL: https://scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html (visited on 08/12/2024).

[89] Miran Seok et al. *Korean Semantic Role Labeling Using Korean PropBank Frame Files*. Pages: 87. Dec. 2016. DOI: 10.14257/astl.2016.142.15.

[90] Ali Shafti et al. *The Response Shift Paradigm to Quantify Human Trust in AI Recommendations*. arXiv:2202.08979 [cs]. Feb. 2022. DOI: 10.48550/arXiv.2202.08979. URL: http://arxiv.org/abs/2202.08979 (visited on 12/23/2023).

[91] Noam Shazeer. *GLU Variants Improve Transformer*. en. arXiv:2002.05202 [cs, stat]. Feb. 2020. URL: http://arxiv.org/abs/2002.05202 (visited on 08/12/2024).

[92] Chereen Shurafa, Kareem Darwish, and Wajdi Zaghouani. "Political Framing: US COVID19 Blame Game". In: *Social Informatics 12th International Conference*. Springer, Cham, 2020, pp. 333–351. DOI: 10.1007/978-3-030-60975-7âĂŰ25. URL: https://link.springer.com/chapter/10.1007/978-3-030-60975-7_25.

[93] Yanchuan Sim et al. "Measuring Ideological Proportions in Political Speeches". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (2013), pp. 91–101. URL: https://aclanthology.org/D13-1010/.

[94] Karen Sullivan. "Three levels of framing". eng. In: *Wiley Interdisciplinary Reviews. Cognitive Science* 14.5 (2023), e1651. ISSN: 1939-5086. DOI: 10.1002/wcs.1651.

[95] Chris Sweeney and Maryam Najafian. "A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 1662–1667. DOI: 10.18653/v1/P19-1162.

[96] Joan C Timoneda and Sebastian Vallejo Vera. "BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformers Models in Political Science Text". en. In: (Feb. 2024).

[97] Tiya Vaj. *Dynamic mask for RoBERTa VS static mask for BERT*. en. Sept. 2023. URL: https://vtiya.medium.com/dynamic-mask-for-roberta-vs-static-mask-for-bert-c997edc9a939 (visited on 04/25/2024).

[98] Ashish Vaswani et al. *Attention Is All You Need*. arXiv:1706.03762 [cs]. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 04/27/2024).

[99] Jesse Vig. "BERTVIZ: A TOOL FOR VISUALIZING MULTI-HEAD SELF-ATTENTION IN THE BERT MODEL". en. In: (2019).

[100] Jiahao Wang, Ning Ma, and Fucheng Wan. "Research on Semantic Role Labeling based on Dependency Parsing". In: *2022 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*. Aug. 2022, pp. 1419–1422. DOI: 10.1109/AEECA55500.2022.9918893. URL: https://ieeexplore.ieee.org/document/9918893 (visited on 12/26/2023).

[101] WordLift. *Frame semantics*. en-US. 2024. URL: https://wordlift.io/blog/en/entity/frame-semantics/ (visited on 07/28/2024).

[102] Ben Wu et al. "SheffieldVeraAI at SemEval-2023 Task 3: Mono and Multilingual Approaches for News Genre, Topic and Persuasion Technique Classification". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 1995–2008. DOI: 10.18653/v1/2023.semeval-1.275. URL: https://aclanthology.org/2023.semeval-1.275 (visited on 11/21/2023).

[103] Qi Yu. "Towards a More In-Depth Detection of Political Framing". In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by Stefania Degaetano-Ortlieb et al. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 162–174. DOI: 10.18653/v1/2023.latechclfl-1.18. URL: https://aclanthology.org/2023.latechclfl-1.18 (visited on 11/23/2023).

[104] Ke Zhang, Peidong Xu, and Jun Zhang. "Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control". In: *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)* (Oct. 2020). Conference Name: 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2) ISBN: 9781728196060 Place: Wuhan, China Publisher: IEEE, pp. 711–716. DOI: 10.1109/EI250167.2020.9347147. URL: https://ieeexplore.ieee.org/document/9347147/ (visited on 12/23/2023).

[105] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. "Classifying the Political Leaning of News Articles and Users from User Votes". In: *Proceedings of the International AAAI Conference on Web and Social Media* 5.1 (2011), pp. 417–424. ISSN: 2334-0770. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14108.

A.

Statistical Analysis of the Datasets

Our research employs two primary datasets for training the SLMuSE-DLF and MuSE-DLF models: the Media Frames Corpus (MFC) [14] and the SemEval-2023 dataset [78]. Both datasets use the same frame annotations from the Framing Codebook by Boydstun and Gross [12], which allows us to have a consistent analytical framework for our study. The descriptions of each frame used in the annotation process are shown in Table A.1.

Subsequent chapters will present a comprehensive examination of the Media Frames Corpus (MFC) and the SemEval 2023 Dataset.

Frame	**Description**
Economic	Costs, benefits, or other financial implications.
Capacity and Resources	Availability of physical, human, or financial resources, and capacity of current systems.
Morality	Religious or ethical implications.
Fairness and Equality	Balance or distribution of rights, responsibilities, and resources.
Legality, Constitutionality, Jurisdiction	Rights, freedoms, and authority of individuals, corporations, and government.
Policy Prescription and Evaluation	Discussion of specific policies aimed at addressing problems.
Crime and Punishment	Effectiveness and implications of laws and their enforcement.
Security and Defence	Threats to the welfare of the individual, community, or nation.
Health and Safety	Health care, sanitation, public safety.
Quality of Life	Threats and opportunities for the individual's wealth, happiness, and well-being.
Cultural Identity	Traditions, customs, or values of a social group in relation to a policy issue.
Public Sentiment	Attitudes and opinions of the general public, including polling and demographics.
Political	Considerations related to politics and politicians, including lobbying, elections, and attempts to sway voters.
External Regulation and Reputation	International reputation or foreign policy of the U.S.

Table A.1.: *Overview of Different Frames and Their Descriptions from Boydstun and Gross [12]*

A.1. Media Frames Corpus (MFC) Dataset

This section provides a comprehensive statistical analysis of the Media Frames Corpus dataset, focusing on the distribution of primary frames and tones, the classification of articles as irrelevant, and key textual characteristics.

A.1.1. Primary Frame Distribution

Table A.2 presents an overview of the primary frame distribution within the dataset, including the absolute numbers and corresponding percentages for each frame category. It is crucial to note that only articles marked as relevant are considered in this analysis. After excluding the irrelevant articles from the dataset, some articles may not have an assigned document-level frame, which are listed under *None* in the table.

Primary Frame	**Absolute**	**Relative (%)**
Political	963	15.79
Legality, Constitutionality, Jurisdiction	955	15.66
Crime and Punishment	794	13.02
Cultural Identity	550	9.02
Policy Prescription and Evaluation	471	7.73
Economic	412	6.76
Quality of Life	408	6.69
Security and Defense	283	4.64
Public Sentiment	243	3.99
Health and Safety	236	3.87
Capacity and Resources	210	3.44
None	204	3.35
Fairness and Equality	155	2.54
External Regulation and Reputation	128	2.10
Morality	76	1.25
Other	9	0.15

Table A.2.: *Overview of main frames in the Media Frames Corpus. This table presents the total number and percentage of each frame, pointing out the dominance of Political and Legality frames, with Other and Morality frames being the least frequent. Only articles marked as relevant are included in this analysis; articles labeled as relevant but lacking a frame are categorized under None.*

A.1.2. Primary Tone Distribution

Here is the paraphrased text: The main tone distribution, illustrated in Table A.3, categorizes articles according to their stance on the subject. The articles are classified as either pro-immigration, neutral on immigration, or anti-immigration. For instance, articles labeled as anti-immigration utilize framing techniques to articulate their resistance to immigration.

Primary Tone	Absolute	Relative (%)
Pro	2718	44.57
Neutral	1708	28.01
Anti	1671	27.40

Table A.3.: *Distribution of primary tones in the Media Frames Corpus dataset. The table presents the absolute count and relative percentage for each tone category (Pro, Neutral, Anti), revealing the overall sentiment balance in the corpus. Only articles marked as relevant are included in this analysis.*

A.1.3. Irrelevant Articles

Articles classified as irrelevant are either non-contributory to the framing analysis or do not pertain to the issues discussed. The distribution is shown in Table A.4. The annotation was done by the Media Frames Corpus dataset creators.

Irrelevant	Absolute	Relative (%)
relevant	6097	90.92
irrelevant	609	9.08

Table A.4.: *Overview of relevant and irrelevant articles within the Media Frames Corpus dataset. The table shows the total number and relative proportion for each category, highlighting the dataset's emphasis on relevant content. Only articles marked as relevant are included in this analysis.*

A.1.4. Textual Characteristics

Textual characteristics such as the length of texts, the number of sentences, and the average number of words per sentence are crucial to understanding the depth of content analysis. These metrics are summarized in Table A.5.

Statistic	Number of Sentences	Number of Words
Mean	12.59	20.00
Standard Deviation	4.43	12.10
Minimum	1	1
Maximum	108	166
25% Percentile	10	10
50% Percentile (Median)	12	19
75% Percentile	14	28
95% Percentile	18	41
99% Percentile	24	52

Table A.5.: *Evaluation of textual properties within the Media Frames Corpus. The table displays essential metrics for the count of sentences and words per article, encompassing central tendency indicators (mean, median) and variability measures (standard deviation, percentiles). Only articles marked as relevant are included in this analysis.*

A.1.5. Articles Distribution

The upcoming sections illustrate diverse distributions of the articles, specifically focusing on articles per source and articles per year.

Articles Per Source

Figure A.1 depicts the distribution of articles among different sources. This illustration helps in understanding the input of various news channels to the dataset. The news sources *new york times* and *washington post* notably have the highest number of articles included in the MFC dataset.

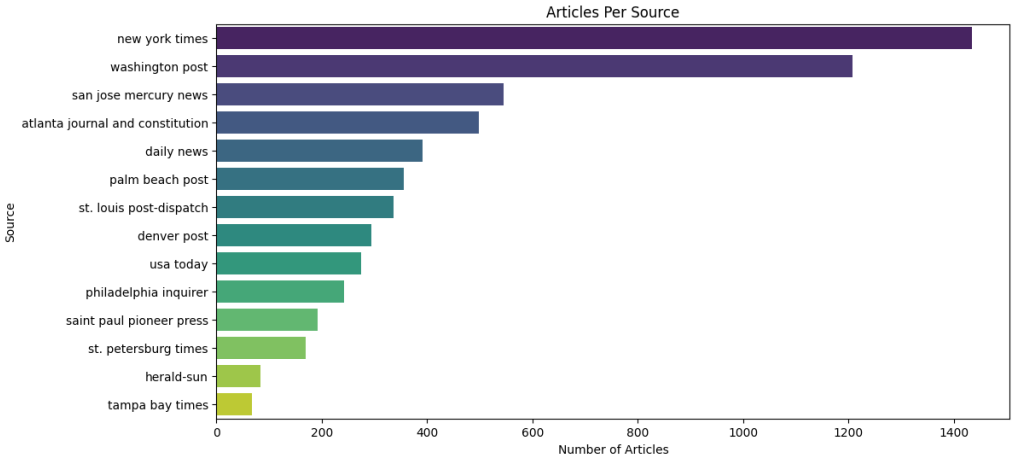

Figure A.1.: *Evaluation of the article count distribution in relation to their publishing sources. Greater values indicate a larger quantity of articles issued by that source. Only articles marked as relevant are included in this analysis.*

Articles Per Year

Figure A.2 shows the yearly distribution of articles, offering a perspective on the publication timeline. The dataset's earliest articles date back to 1969, with the latest ones published in 2017. It is apparent that the majority of articles in the MFC are from the 2000-2017 period.

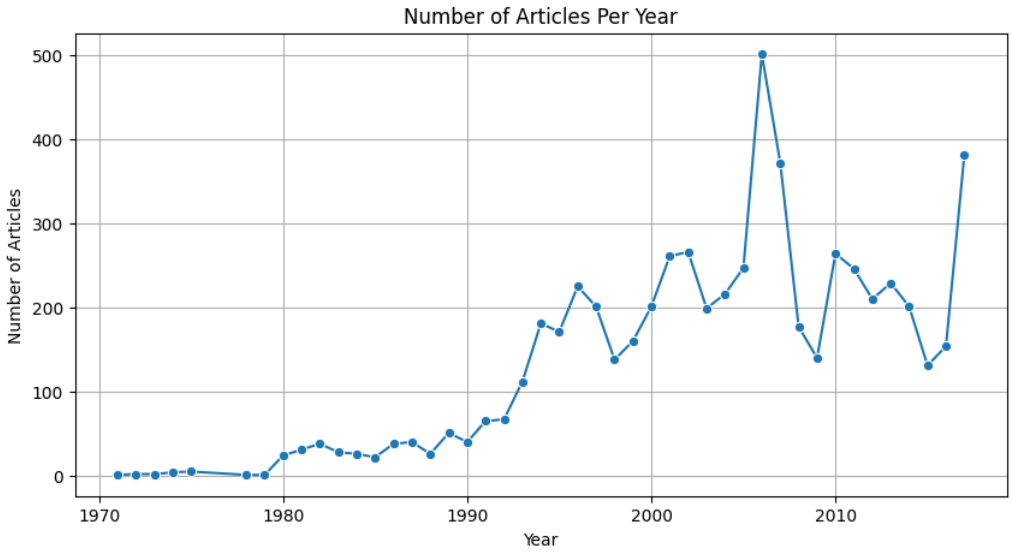

Figure A.2.: *Analysis of the yearly publication count of articles. Only articles marked as relevant are included in this analysis.*

A.1.6. Primary Frame and Tone Analysis

Examination of the distribution of various document frames across three distinct tones relating to the policy issue (pro-immigration, anti-immigration, and neutral-immigration).

Primary Frame and Tone Co-Occurence

The relationship between Primary Frame and tone is depicted through a graph to analyze how various tones are distributed among the frames. The graph indicates that 57% of the articles annotated with *Crime and Punishment* exhibit an *anti-immigration* tone. Conversely, articles tagged with *Morality* or *Quality of Life* predominantly display a *pro-immigration* tone with 84% and 83%.

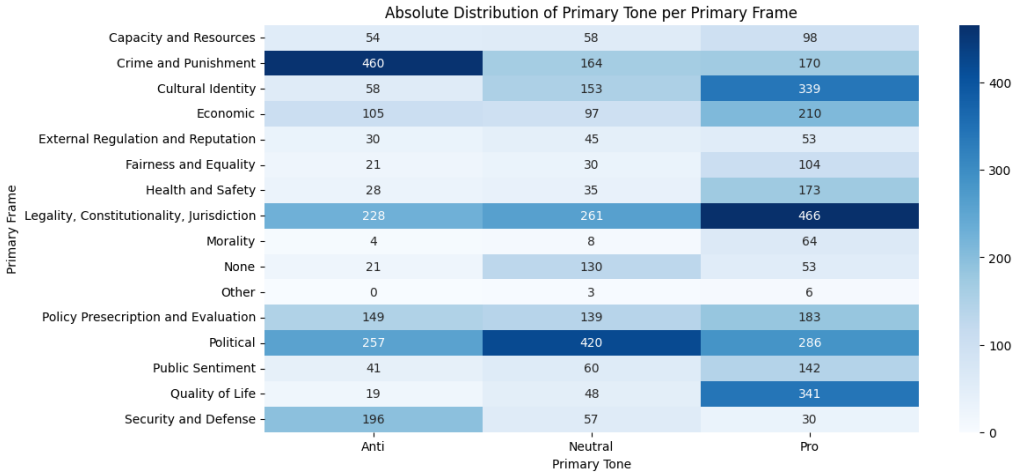

Figure A.3.: *Analysis of the usage of tones in article document frames. Darker shades indicate a higher quantity of articles with that specific tone. Only articles marked as relevant are included in this analysis.*

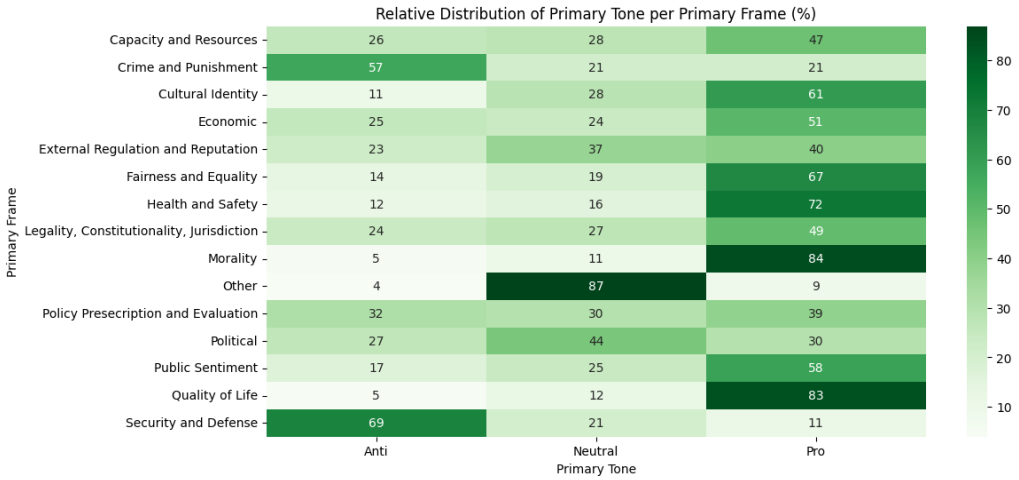

Figure A.4.: *Analysis of the usage of tones in article document frames. Darker shades represent a greater relative occurrence of articles featuring that specific tone. The values are normalized across rows. Only articles marked as relevant are included in this analysis.*

Source Influence on Framing

Examination of the distribution of articles by various sources across different document frames and news outlets.

Primary Frame Distribution by Source Frames such as *Crime and Punishment*, *Legality, Constitutionality, Jurisdiction*, and *Political* are overly prevalent in comparison to others, as indicated by the overall frame distribution shown in table A.2. In particular, articles with a frame

Political are prominently displayed, and *Tampa Bay Times* assigns a frame *Political* to more than 39% of its analyzed articles.

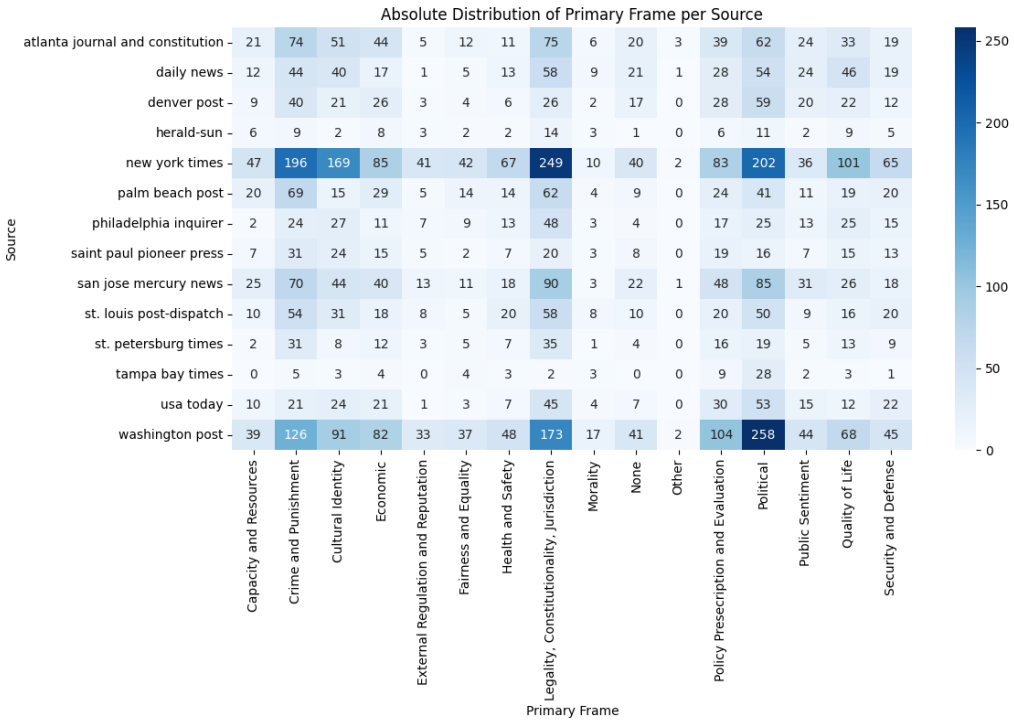

Figure A.5.: *Absolute distribution of document frames across publishing sources. Darker shades indicate a greater frequency of source-primary frame co-occurrence. Only articles marked as relevant are included in this analysis.*

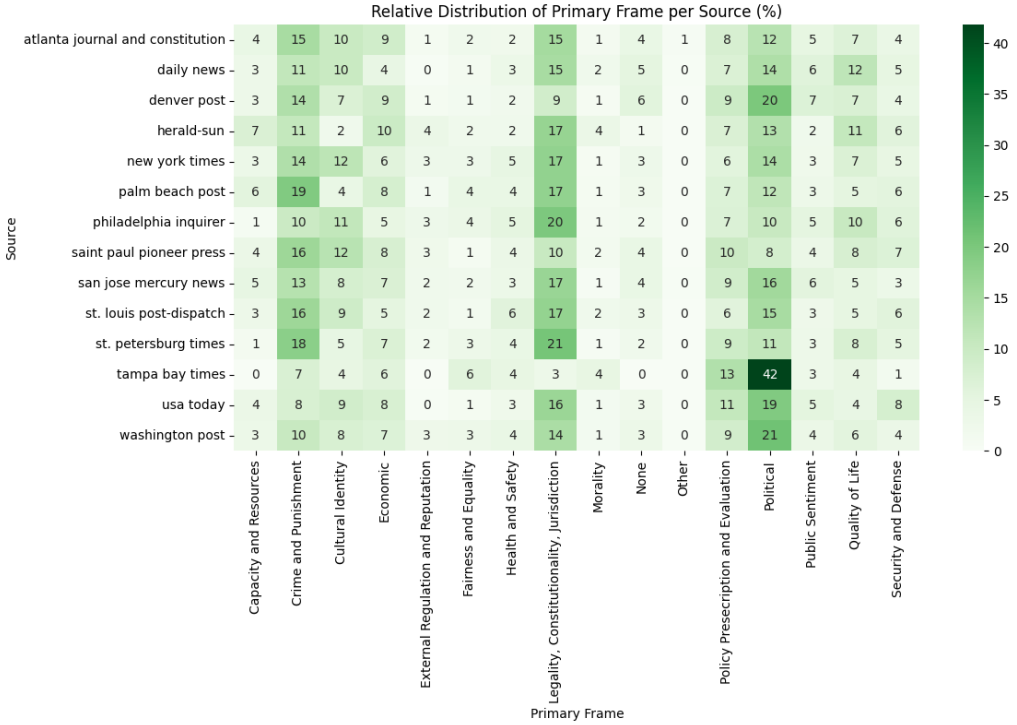

Figure A.6.: *Relative frequency distribution of document frames among publishing sources. Darker colors suggest a greater relative occurrence of source-primary frame matches. The percentages are normalized for each row. Only articles marked as relevant are included in this analysis.*

Primary Tone Distribution by Source The widespread dominance of *pro-immigration* tone articles, as shown in table A.3, is also reflected at the level of individual newspaper publishers. Specifically, newspapers like the *Herald Sun*, *Daily News*, and *San Jose Mercury News* have a higher proportion of *pro-immigration* articles compared to the overall trend. Conversely, the *Denver Times* presents more *anti-immigration* tone articles than *pro-immigration* ones. Similarly, the *St. Petersburg Times* shows an equal distribution of 31% *pro-immigration* and 31% *anti-immigration* articles, thus differing from the general distribution pattern.

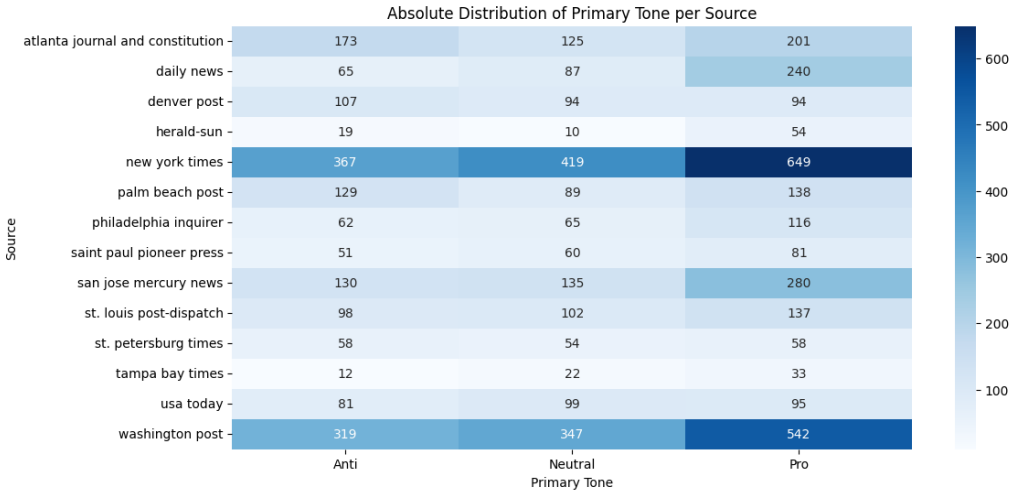

Figure A.7.: *Analysis depicting the absolute distribution of main tones among various sources. Darker colors indicate a greater absolute frequency of co-occurrence. Only articles marked as relevant are included in this analysis.*

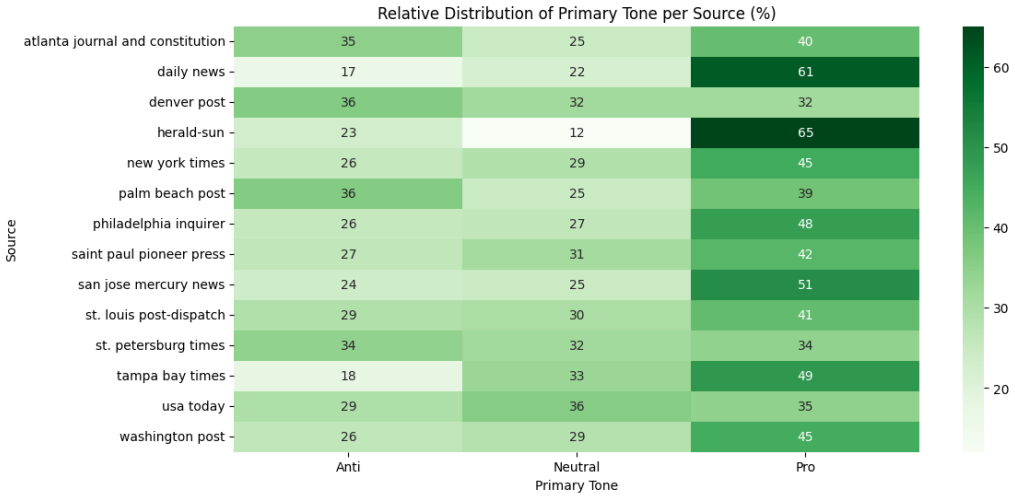

Figure A.8.: *Analysis depicting the relaitve distribution of main tones among various sources. Darker colors indicate a greater relative frequency of co-occurrence. Only articles marked as relevant are included in this analysis.*

A.1.7. Semantic Role Labeling Analysis

The table below (see Table A.6) summarizes the statistical evaluation of the semantic roles per sentence within the dataset. It encompasses a range of statistical metrics, including count, average, standard deviation, minimum, several percentiles, and the maximum count of arguments in sentences.

Statistic	Number of Semantic Roles per Sentence	Number of Words per Semantic Role
Mean	4.29	3.07
Standard Deviation	3.15	5.43
Minimum	0	0
Maximum	41	119
25% Percentile	2	1
50% Percentile (Median)	4	1
75% Percentile	6	3
95% Percentile	10	14
99% Percentile	13	28

Table A.6.: *Statistical Overview of Semantic Roles per Sentence and Word Count per Semantic Role. The columns depicting the number of semantic roles per sentence present different statistical metrics indicating the occurrence of semantic roles per sentence, whereas the number of words per semantic role emphasizes the word count within these semantic role spans. Only articles marked as relevant are included in this analysis.*

A.2. SemEval 2023 Dataset

This section provides a comprehensive statistical analysis of the SemEval 2023 dataset, focusing on the distribution of primary frames and tones, the classification of articles as irrelevant, and key textual characteristics. The frame annotations of the dataset at the document level were performed by around 40 expert annotators following the guidelines of The Framing Codebook [12]. A comprehensive statistic of the dataset is presented in Table A.7, which includes the number of documents, characters, and average frames per language.

Language	Documents	Characters	Avg. Frames
English	446	2,431K	3.7
French	158	737K	3.0
German	132	581K	4.3
Italian	227	927K	3.8
Polish	145	765K	5.0
Russian	143	590K	2.5

Table A.7.: *Statistics about the data for different languages: total number of articles, characters, and average number of frames per document.*

A.2.1. Frames Distribution

Table A.8 presents an overview of the dataset's frame distribution, highlighting both the absolute counts and the relative proportions for each frame category. This table underscores the frames that

are most and least prevalent, with *Political* and *Crime and Punishment* being the most frequent, while *Public Opinion* is the least frequent.

Frames	**Absolute**	**Relative (%)**
Economic	28	1.73%
Capacity and resources	29	1.80%
Morality	203	12.57%
Fairness and equality	115	7.12%
Legality, constitutionality, and jurisprudence	203	12.57%
Policy prescription and evaluation	65	4.02%
Crime and punishment	227	14.06%
Security and defense	187	11.58%
Health and safety	61	3.78%
Quality of life	87	5.39%
Cultural identity	31	1.92%
Public opinion	23	1.42%
Political	235	14.55%
External regulation and reputation	121	7.49%
Total	**1615**	**100%**

Table A.8.: *Frame distribution of English articles in the SemEval dataset represented in both absolute and relative terms.*

A.2.2. Type Distribution

Table A.9 shows the distribution of article types, demonstrating the categorization of articles. Articles are classified as either *opinion*, *reporting*, or *satire*.

Type	**Absolute**	**Relative (%)**
Opinion	382	88.22
Reporting	41	9.47
Satire	10	2.31

Table A.9.: *Distribution of primary tones in the SemEval dataset. The table presents the absolute count and relative percentage for each article type category (Opinion, Reporting and Satire), revealing the article type balance in the corpus*

A.2.3. Textual Characteristics

Attributes of the text, including text length, sentence count, and the mean number of words per sentence, are essential for comprehending the extent of content analysis. These measurements are detailed in Table A.10.

Statistic	Number of Sentences	Number of Words
Mean	39.01	23.29
Standard Deviation	35.55	17.65
Minimum	4	1
Maximum	408	184
25% Percentile	19	11
50% Percentile (Median)	30	19
75% Percentile	44	31
85% Percentile	57	38
90% Percentile	69	44
95% Percentile	108	56
99% Percentile	166	90

Table A.10.: *Evaluation of textual properties within the SemEval dataset. The table displays essential metrics for the count of sentences and words per article, encompassing central tendency indicators (mean, median) and variability measures (standard deviation, percentiles).*

A.2.4. Co-occurrence of Frames and Types

This part explores the simultaneous occurrence of frames with various content types (opinion, reporting, satire). Grasping these patterns uncovers the ways different frames are employed across content types, offering a detailed perspective on frame usage in media and other settings. The examination features both absolute and relative value heatmaps.

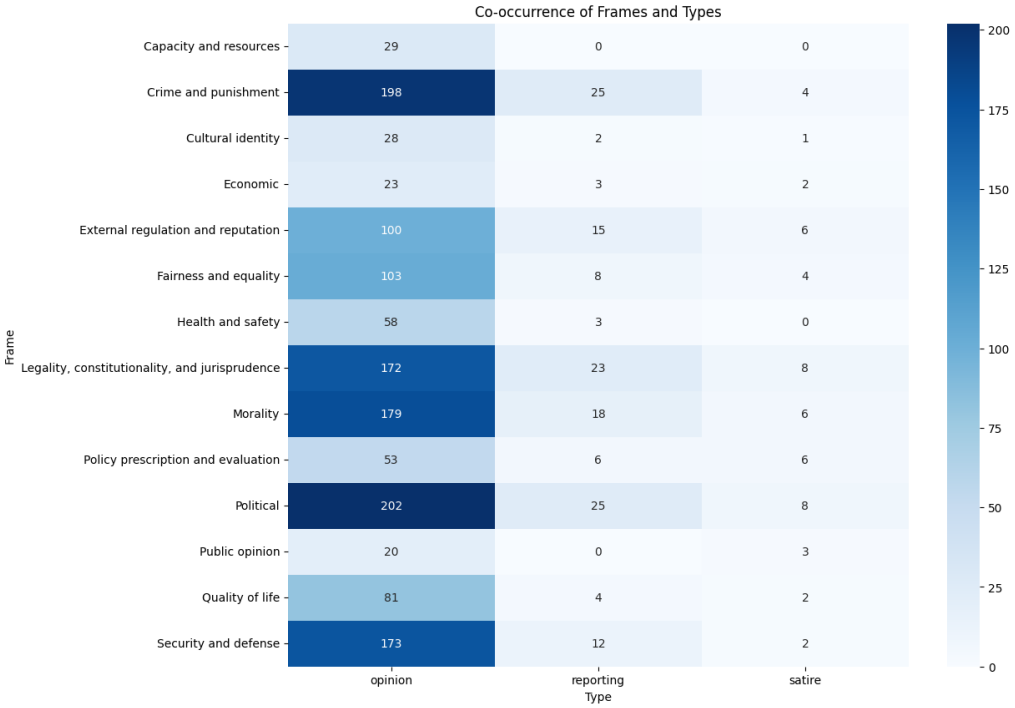

Figure A.9.: *Heatmap showing the absolute co-occurrence of frames across various types (opinion, reporting, satire). Darker shades represent higher frequencies, with significant co-occurrences such as 'Crime and Punishment' in opinion articles and 'Political' frames appearing in multiple types.*

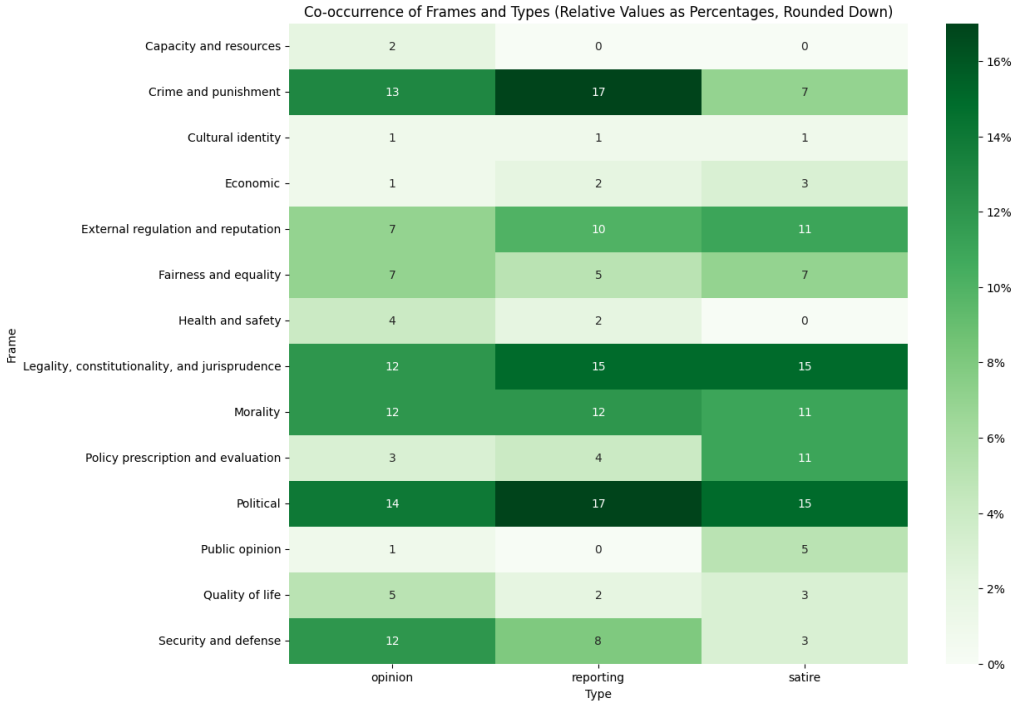

Figure A.10.: *Heatmap illustrating the relative co-occurrence of various frame types, presented as percentages and rounded down. The diagram emphasizes notable relative frequencies, such as 'Morality' and 'Legality, Constitutionality, and Jurisprudence' in opinion articles and 'Political' frames in news reports.*

A.2.5. Semantic Role Labeling Analysis

The table below (see Table A.11) summarizes the statistical evaluation of the semantic roles per sentence within the dataset. It encompasses a range of statistical metrics, including count, average, standard deviation, minimum, several percentiles, and the maximum count of arguments in sentences.

Statistic	Number of Semantic Roles per Sentence	Number of Words per Semantic Role
Mean	4.95	3.08
Standard Deviation	4.21	6.37
Minimum	0	0
Maximum	42	200
25% Percentile	2	1
50% Percentile (Median)	4	1
75% Percentile	7	2
85% Percentile	9	5
90% Percentile	10	8
95% Percentile	13	14
99% Percentile	20	32

Table A.11.: *Statistical Overview of Semantic Roles per Sentence and Word Count per Semantic Role. The columns depicting the number of semantic roles per sentence present different statistical metrics indicating the occurrence of semantic roles per sentence, whereas the number of words per semantic role emphasizes the word count within these semantic role spans.*

B.

Experiments - Additional Resources

B.1. Tokenization Process and Word-to-Token Ratio

In this research, we use the RoBERTa-base tokenizer, which employs Byte-Pair Encoding (BPE) for breaking down words into subwords [43]. Since BPE can generate multiple tokens from a single word, the word count often differs from the token count. Understanding this relationship is crucial for our model, as we depend on word count to estimate the number of tokens per sentence or for each semantic role. To explore this ratio in our datasets, we analyzed samples from the Media Frames Corpus (MFC) and the SemEval dataset.

We reviewed 5 randomly selected articles from each dataset, focusing solely on relevant articles that had frames assigned for the MFC. The findings are displayed in Table B.1.

Dataset	Article	Avg. Token Word Ratio	Std. Deviation
Media Frames Corpus	1	1.31	0.21
	2	1.31	0.22
	3	1.29	0.13
	4	1.33	0.23
	5	1.29	0.28
SemEval	1	1.28	0.09
	2	1.37	0.22
	3	1.47	0.46
	4	1.26	0.18
	5	1.37	0.64
Media Frames Corpus Average		1.31	0.21
SemEval Average		1.35	0.32

Table B.1.: *Analysis of the Token-to-Word Ratio for five sample articles from the MFC and SemEval datasets*

The analysis reveals that both datasets exhibit similar token-to-word ratios, with the SemEval dataset averaging 1.35 tokens per word and the MFC averaging 1.31 tokens per word.

B.2. Experiment 1

Subsequent chapters offer addtional plots and ressources for experiment 1.

B.2.1. Focal Loss for Multi-Class Classification

Focal Loss, as proposed by Lin et al. [61], is a modification to the traditional cross-entropy loss aimed at tackling class imbalance issues in classification problems. This part elaborates on the Focal Loss implementation utilized in our research. We use the unoffocial PyTorch implementation from Hassan [36].

Focal Loss for a single sample is defined as:

$$\mathrm{FL}(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t) \tag{B.1}$$

where p_t is the model's estimated probability for the true class t, α_t is the class balancing factor for class t and γ is the focusing parameter ($\gamma \geq 0$)

For multi-class problems, p_t is computed using the softmax function:

$$p_t = \frac{\exp(x_t)}{\sum_{j=1}^{C} \exp(x_j)} \tag{B.2}$$

where x_t is the model's output (logit) for the true class t, and C is the total number of classes.

Key Properties:

1. The factor $(1-p_t)^{\gamma}$ decreases the loss for examples that are classified correctly.
2. α_t mitigates class imbalance by modifying the weight assigned to each class.
3. γ determines how quickly the weight of easy examples is reduced.

For an in-depth examination of the Focal Loss function and its characteristics, consult the foundational paper by Lin et al. [61] or the Code implementation [36].

B.2.2. Semantic Axis analysis

The subsequent section presents additional Semantic Axis plots related to the Semantic Axis analysis in 6.2.2.

appendix:b

appendix:b

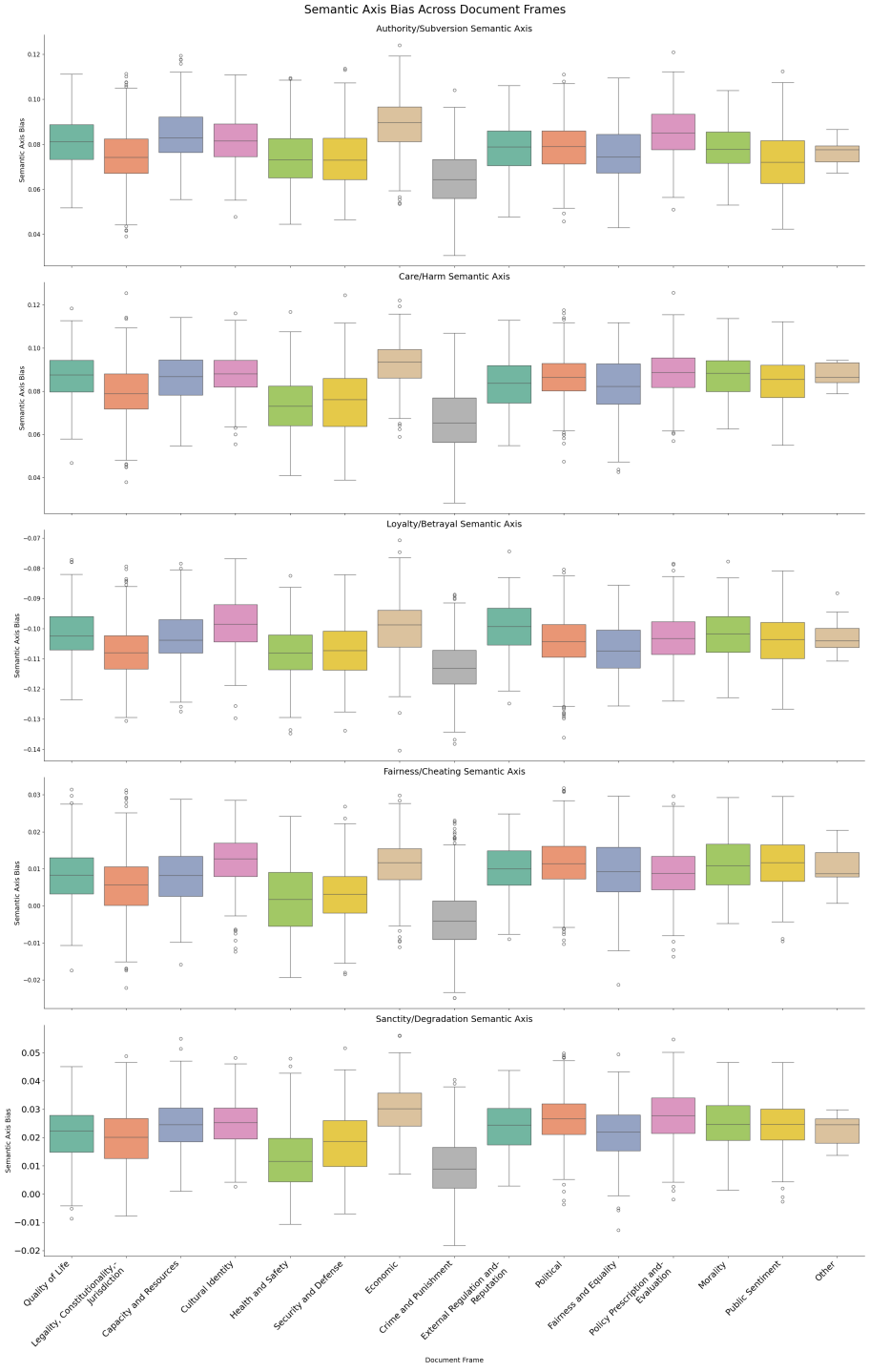

Figure B.1.: *Semantic Axis Bias Across Document Frames. This figure shows box plots comparing semantic axis biases for different document frames across multiple semantic dimensions.*

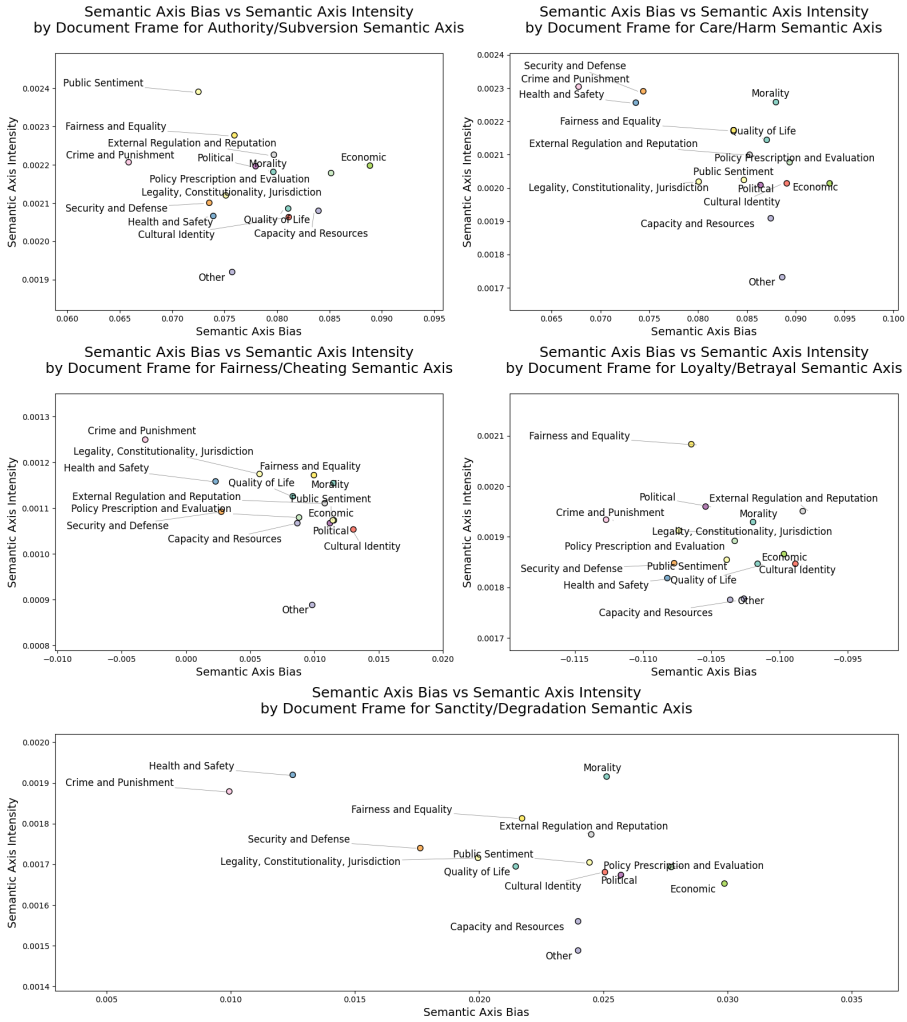

Figure B.2.: *Semantic Axis Bias vs Semantic Axis Intensity by Document Frame. This figure presents scatter plots showing the relationship between semantic axis bias and intensity for different document frames across multiple semantic dimensions.*

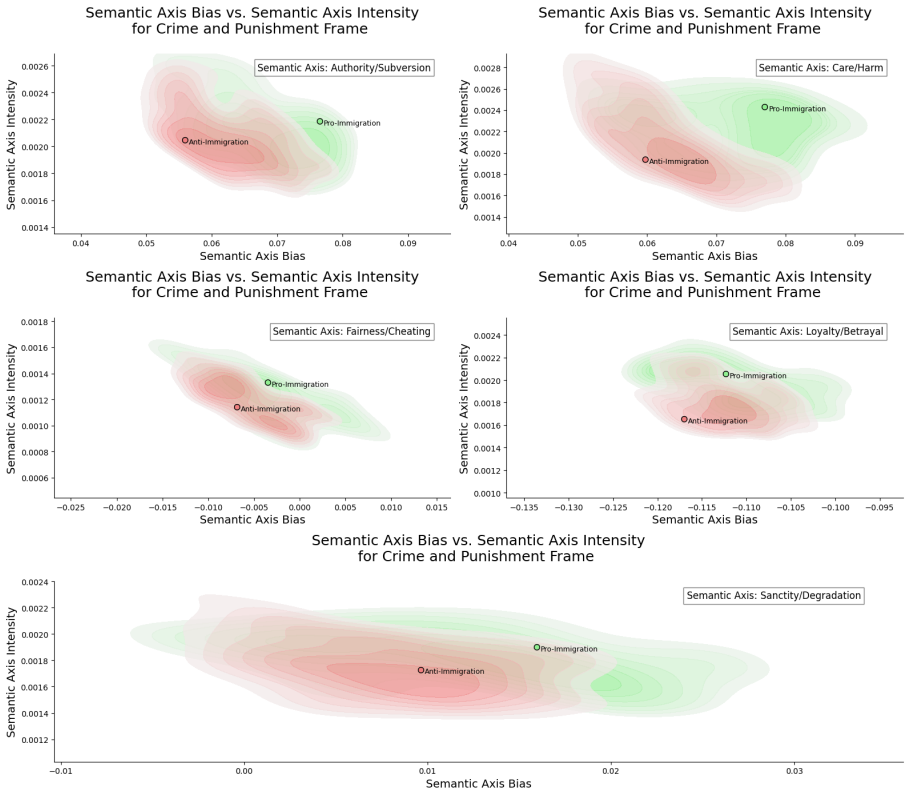

Figure B.3.: *Semantic Axis Bias vs. Semantic Axis Intensity for Crime and Punishment Frame. This figure displays scatter plots with density contours, illustrating the relationship between semantic axis bias and intensity specifically for the Crime and Punishment frame across various semantic dimensions.*

Figure B.4.: *Bias shifts in semantic axes for all frames, tones, and semantic axes. This figure shows the top 5 words with the greatest absolute shifts in semantic axis bias for all frames, tones, and semantic axes, providing insights into how different words influence the framing towards specific poles. (Limited view. Full version available online)*

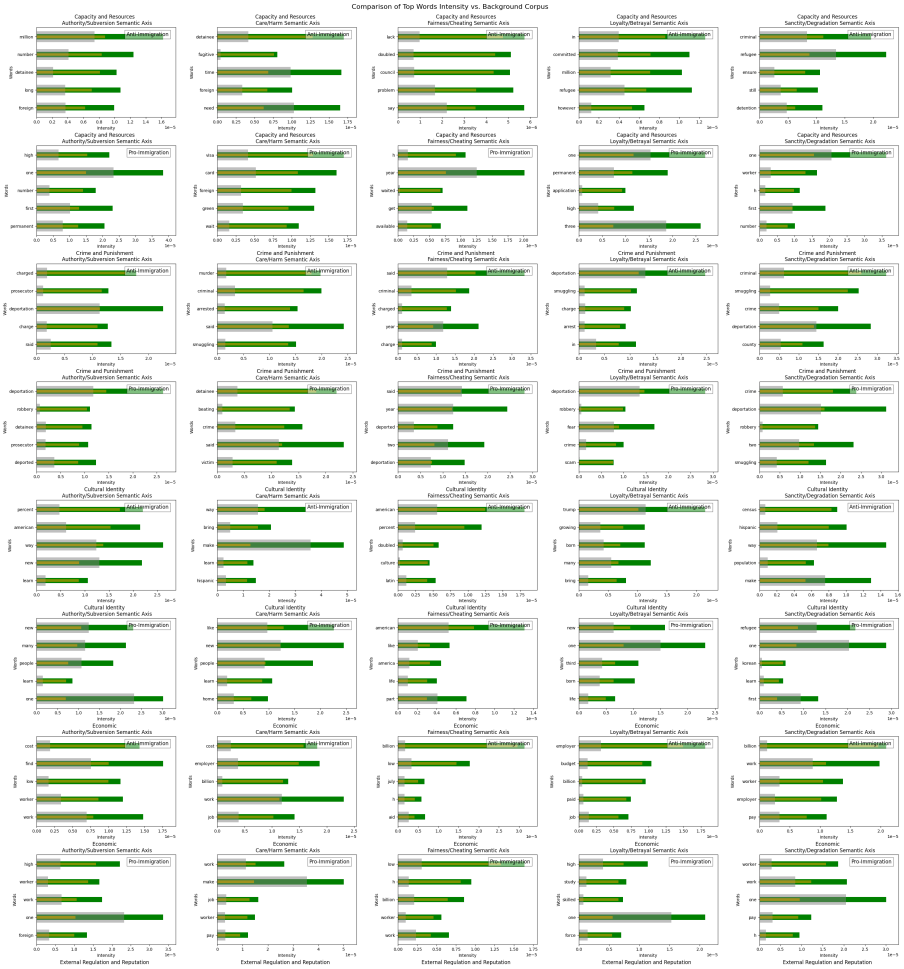

Figure B.5.: *Intensity shift in semantic axes for all frames, tones, and semantic axes. This figure highlights the top-5 words exhibiting the greatest intensity changes across all frames, tones, and semantic axes, providing an understanding of the most frequently used words within these semantic axes for each frame and tone. (Limited view. Full version available online)*

B.2.3. Full Dataset-level Semantic Role Analysis

Table B.2 presents the complete dataset-level semantic role analysis for all frames.

B.3. Experiment 2

B.3.1. Asymmetric Loss Function

The Asymmetric Loss (ASL) function, introduced by Ben-Baruch et al. [9], offers a new method to tackle class imbalance in multi-label classification tasks. This section focuses on how ASL was implemented in our study, based on the math from the paper [9].

The ASL for a single sample is defined as:

$$ASL = \begin{cases} (1-p)^{\gamma_+} \log(p) & \text{if } y = 1 \\ (p_m)^{\gamma_-} \log(1-p_m) & \text{if } y = 0 \end{cases} \tag{B.3}$$

where y is the ground truth label (1 for positive, 0 for negative), p is the predicted probability, $p_m = \max(p-m, 0)$, γ_+ and γ_- are focusing parameters for positive and negative samples respectively, and m is the probability margin (clip).

The ASL function introduces asymmetry through two key mechanisms:

1. **Asymmetric Focusing:** Distinct focusing parameters (γ_+ and γ_-) for positive and negative samples enable separate handling of false positives and false negatives. This aids in balancing the impact of infrequent positive samples against the numerous negative samples.

2. **Probability Shifting:** The margin m adjusts the probabilities of negative samples, thereby removing the simplest negative samples. This allows the training process to concentrate on more difficult examples.

In our approach, the loss function was additionally adjusted with class-specific weights (α) to mitigate class imbalance. Hence, the ultimate weighted ASL integrates class weighting, uneven handling of positive and negative samples, and probability clipping.

Key Properties:

1. The term $(1-p_m)^{\gamma_+}$ for positive instances and $(p_m)^{\gamma_-}$ for negative instances adjusts the loss, lowering it for correctly classified samples and raising it for those that are incorrectly classified.

2. The differing focusing parameters (γ_+ and γ_-) enable separate handling of false positives versus false negatives, thus tackling class imbalance.

3. The probability threshold m eliminates overly simple negative samples, enabling the model to concentrate on more difficult instances.

Frame	Predicate	Agent	Theme
Quality of Life	flee, introduced, created, sponsored, gave, end, work, lost, start, received	family, father, refugee, young, four, mr, parent, mother, woman, legal	family, child, wife, school, better, life, mother, home, practice, ice
Other	help, proposed, leave, turned, led, called, enter, keep, live, call	latino, jose, another, bush, order, state, immigrant, gop, president, man	demand, republican, number, welfare, federal, result, immigrant, immigration, largest, relative
Crime and Punishment	convicted, identify, target, threatened, served, charged, involved, arrested, thought, face	agent, police, team, authority, patrol, custom, enforcement, border, chief, official	gang, car, individual, released, two, detainee, operation, drug, man, suspect
Economic	hire, intended, hiring, pay, consider, provided, tell, work, fill, estimated	bank, grower, employer, company, job, worker, college, study, business, association	labor, worker, tuition, economic, job, cost, business, farm, thousand, percent
External Regulation and Reputation	apply, worked, put, needed, seen, done, received, leave, return, made	atlanta, hispanic, public, san, immigrant, two, house, father, people, immigration	domestic, couple, construction, guard, made, immigrant, immigration, district, island, approach
Policy Prescription and Evaluation	report, left, charged, stop, mean, called, asked, include, took, got	provision, officer, illegal, country, resident, immigrant, district, legal, immigration, trump	system, benefit, immigrant, law, immigration, allow, legal, check, change, illegal
Fairness and Equality	died, came, deported, issued, set, face, bring, called, continue, living	chairman, security, law, two, several, president, parent, immigrant, homeland, official	road, black, trying, site, similar, employee, skilled, provision, since, immigrant
Legality, Constitutionality, Jurisdiction	ruled, refused, denied, bar, undocumented, require, begin, accept, apply, prove	judge, court, barack, law, civil, living, state, illegally, appeal, order	hearing, basis, removal, order, notice, ruling, commission, request, lived, appeal
Security and Defense	help, arrested, passed, feel, signed, forced, held, began, pay, deport	human, city, mr, president, immigrant, worker, report, trump, undocumented, administration	border, hope, whether, way, immigration, security, immigrant, lawsuit, strong, united
Capacity and Resources	speak, coming, reported, convicted, allows, came, saying, run, given, detained	child, yearold, student, immigrant, year, federal, people, report, office, state	campaign, washington, door, noncitizen, immigration, back, immigrant, tax, grant, college
Morality	need, failed, include, let, remain, took, working, came, using, provide	trump, official, woman, congress, state, lawyer, republican, immigration, community, mr	detail, whose, along, housing, cause, federal, immigrant, undocumented, government, enough
Political	gone, build, decided, approved, sponsored, made, show, prevent, support, related	fox, party, republican, romney, senate, white, rep, senator, president, gop	politician, ad, reform, pressure, democratic, republican, immigration, bill, bipartisan, tough
Public Sentiment	heard, prove, enforce, signed, leave, failed, saying, offer, set, estimated	garcia, haitian, immigrant, people, government, democrat, official, immigration, national, measure	laborer, away, today, secure, rally, florida, immigrant, illegal, entry, benefit
Cultural Identity	lead, growing, added, wrote, report, held, build, released, came, received	community, population, city, york, study, school, immigrant, san, group, authority	culture, decade, census, hispanic, growth, population, value, english, immigrant, american
Health and Safety	entered, released, support, arrested, took, went, announced, working, face, found	spokesman, judge, recent, system, mr, parent, worker, secretary, agent, immigration	entered, cuban, town, dream, illegal, protect, immigrant, foreignborn, flight, yearold

Table B.2.: *Full Dataset-level Semantic Role Analysis for the MFC dataset*

4. The extra class weights α help address class imbalance by adjusting the importance given to each class.

For an in-depth examination of the ASL function and its characteristics, consult the foundational paper by Ben-Baruch et al. [9].

B.3.2. Semantic Axis analysis

The subsequent section presents additional Semantic Axis plots related to the Semantic Axis analysis in 6.3.2.

appendix:b

appendix:b

B.3.3. Full Dataset-level Semantic Role Analysis

Table B.3 presents the complete dataset-level semantic role analysis for all frames.

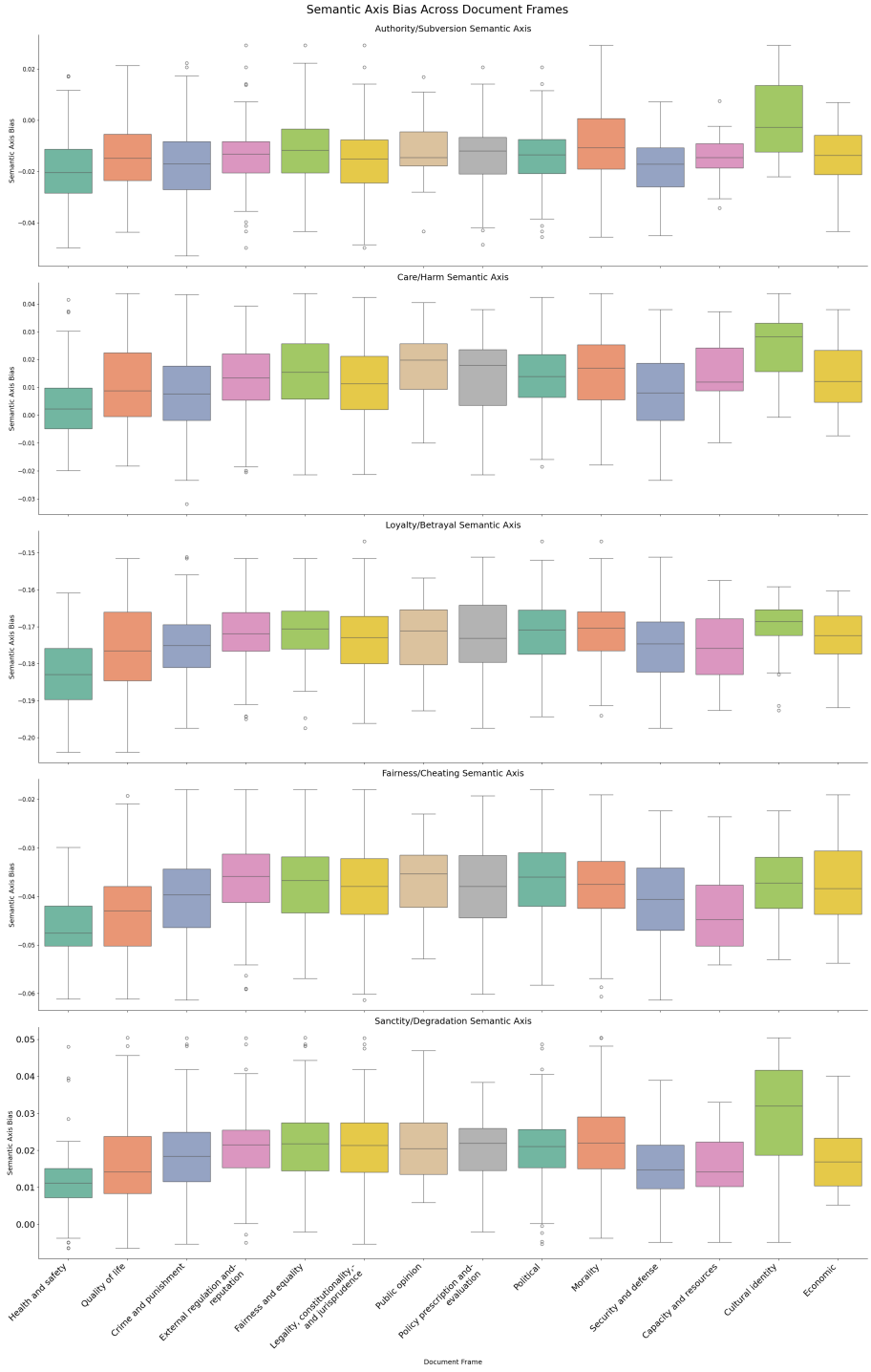

Figure B.6.: *Semantic Axis Bias Across Document Frames. This figure shows box plots comparing semantic axis biases for different document frames across multiple semantic dimensions.*

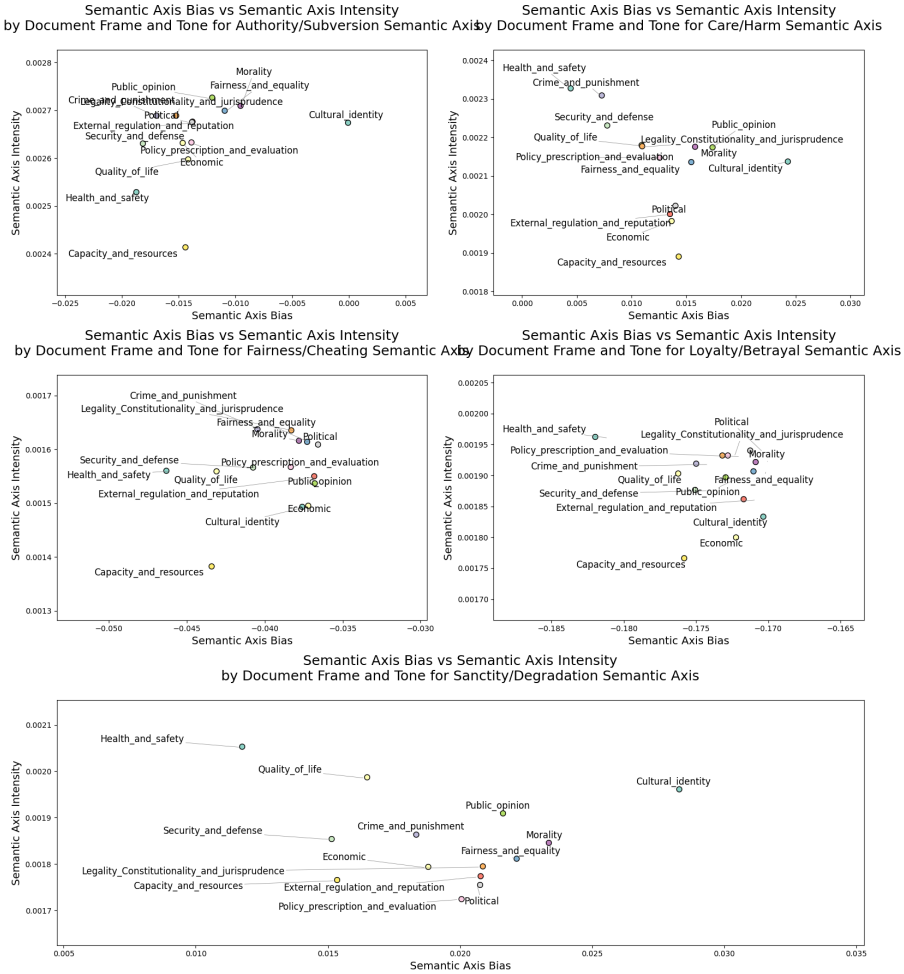

Figure B.7.: *Semantic Axis Bias vs Semantic Axis Intensity by Document Frame. This figure presents scatter plots showing the relationship between semantic axis bias and intensity for different document frames across multiple semantic dimensions.*

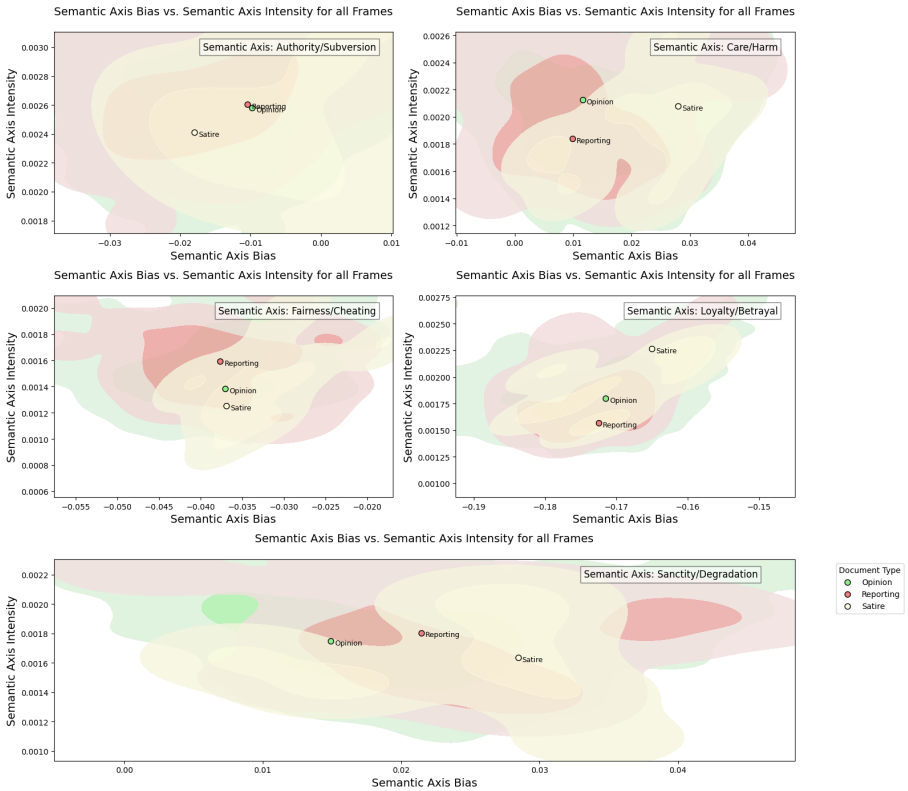

Figure B.8.: *Semantic Axis Bias vs. Semantic Axis Intensity for Crime and Punishment Frame. This figure displays scatter plots with density contours, illustrating the relationship between semantic axis bias and intensity specifically for the Crime and Punishment frame across various semantic dimensions.*

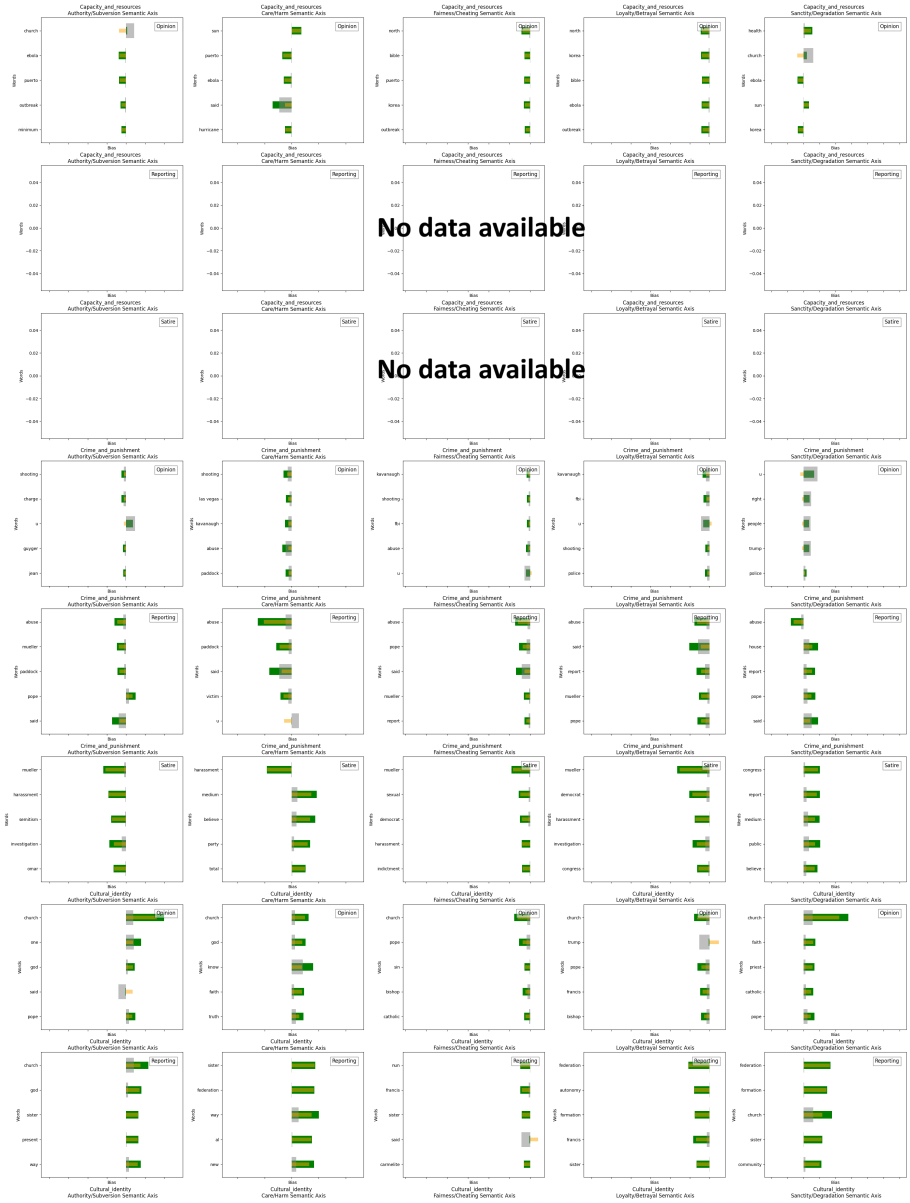

Figure B.9.: *Bias shifts in semantic axes for all frames, types, and semantic axes. This figure shows the top 5 words with the greatest absolute shifts in semantic axis bias for all frames, types, and semantic axes, providing insights into how different words influence the framing towards specific poles. The label "No data available" indicates that no data exists for this particular combination of document frame and tone. For example, there are no articles in the dataset marked as Capacity and Resources that are also labeled as Satire. (Limited view. Full version available online)*

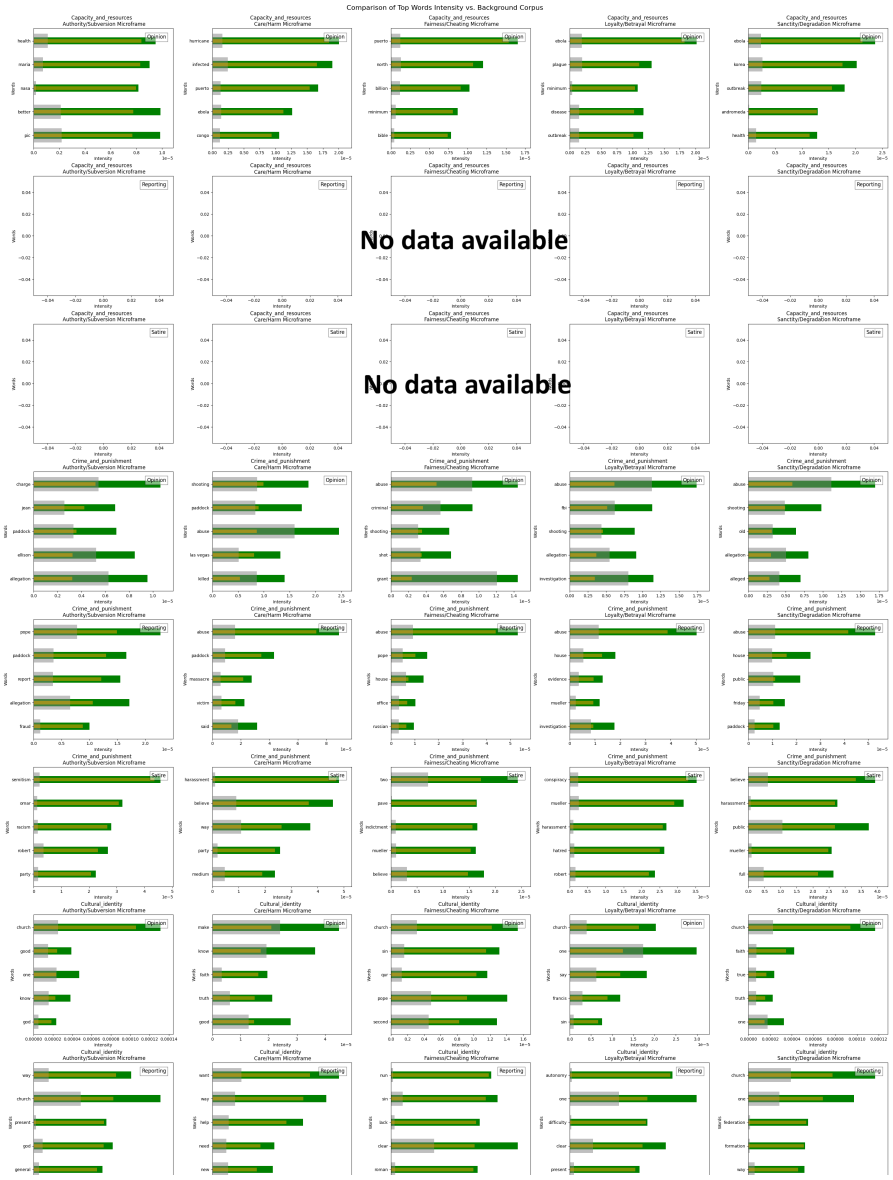

Figure B.10.: *Intensity shift in semantic axes for all frames, type, and semantic categories. This figure highlights the top-5 words exhibiting the greatest intensity changes across all frames, type, and semantic axes, providing an understanding of the most frequently used words within these semantic axes for each frame and tone. The label "No data available" indicates that no data exists for this particular combination of document frame and tone. For example, there are no articles in the dataset marked as Capacity and Resources that are also labeled as Satire. (Limited view. Full version available online)*

Frame	Predicate	agent	theme
Fairness and equality	bring, need, question, implemented, made, vote, added, learn, spoke, disclose	trump, president, debate, cardinal, storm, state, report, china, police, home	muslim, american, people, church, trump, man, kritarch, al, state, gay
Security and defense	arrested, stop, killed, marching, found, missing, shot, entering, died, charged	migrant, government, police, trump, iranian, american, health, officer, uk, official	weapon, gun, compound, military, vega, la, operation, muslim, people, attack
External regulation and reputation	bring, leave, made, fed, secure, left, believe, stop, continuing, insists	trump, iran, president, state, leader, authored, mueller, obama, united, security	trump, brexit, iran, deal, report, russia, eu, president, mueller, haley
Cultural identity	reaffirmed, made, leaving, begin, allowed, report, voting, sue, described, underestimate	trump, steyer, president, rick, duane, francis, surprise, people, news, department	judy, determined, supply, exists, criminal, work, parliamentary, investigation, caravan, brexit
Morality	continues, understand, completing, wrote, saying, refuse, attended, talking, update, left	cardinal, pope, church, francis, priest, bishop, ali, catholic, mller, people	god, bishop, pope, field, farrakhan, jew, church, francis, islam, blessing
Capacity and resources	detailing, sit, ask, seeing, coming, accused, sending, search, made, going	teenager, trump, anybody, violence, official, alien, government, president, department, mp	people, pope, inference, arabic, fatality, opposed, deal, vote, story, successful
Crime and punishment	released, arrested, took, found, left, made, acting, shot, reported, grant	police, attorney, prosecutor, officer, fbi, authority, senator, general, senate, investigator	investigation, report, mueller, evidence, police, kavanaugh, officer, illegal, charge, story
Public opinion	talking, trying, reported, published, revealing, walked, saw, asked, agree, occur	allah, orbiter, outbreak, mayor, white, trump, labour, house, communication, government	people, infiltration, character, pope, credibility, border, bishop, terrorist, thing, vacation
Health and safety	spread, infected, affect, pretend, contain, enter, wanted, reported, replace, quoted	enemy, people, francis, witness, faggioli, trump, adam, pope, house, health	plague, ebola, outbreak, disease, vaccine, madagascar, fever, virus, reported, health
Political	continues, agree, avoid, made, confirmed, completing, got, came, grant, replace	trump, president, house, government, leader, administration, white, obama, iran	trump, deal, report, poll, president, freedom, outpost, mueller, access, russia
Policy prescription and evaluation	promotes, reported, seen, reminds, fill, moved, included, wrote, report, offering	belief, muslim, student, vote, chilean, medium, people, officer, police, fbi	people, ebola, going, arm, minor, report, man, belief, surrounding, urban
Legality Constitutionality and jurisprudence	opt, agree, confirmed, grant, released, left, release, continues, saying, believe	attorney, trump, doj, ford, mueller, kavanaugh, department, democrat, house, law	trump, mueller, kavanaugh, information, report, poll, story, investigation, president, fbi
Quality of life	spread, infected, kill, mean, reported, spreading, experiencing, given, seen, develop	researcher, patient, health, plague, obtained, official, trump, president, people, world	plague, outbreak, disease, ebola, madagascar, symptom, bacteria, vaccine, mtbi, virus
Economic	deployed, throwing, admitted, signing, denounced, threatened, authored, saying, headed, try	air, austrian, police, delay, american, britain, country, trump, campus, hierarchy	party, plate, people, ukraine, trump, country, order, example, catholic, person

Table B.3.: *Full Dataset-level Semantic Role Analysis for the SemEval dataset*

C.

Other

C.1. Extended Moral Foundation Key Words

Moral Foundation	Category	Words List
Authority / Subversion	Vice	protested, rebellion, protesting, revenge, rage, violate, accuse, contempt, intimidation, refused, disruption, riot, launched, prosecuted, beating, protesters, condemned, demanding, disagreed, suspicion, protest, rebels, backed, perpetrators, threatens, terrorists, criticized, hostile, rebel, challenged, questioning, protests, engage, ignore, blocking, failing, defenders, fired, lobbying, decisive
Authority / Subversion	Virtue	recommended, authority, revive, promising, embrace, wise, charity, preventing, strongly, loyal, granted, useful, encouraged, stable, supporters, strengthen, convince, stopping, trust, incentives, ambitious, powerful, celebrated, outgoing, adequate, determination, backing, poised, favor, determined, challenge, approved, accepted, ability, improve, gained, approval, hand, prominent, likes
Care / Harm	Vice	tortured, cruel, harsh, hostility, killed, punishments, attacked, suicide, assassination, hurt, waste, pain, threatening, kill, torture, murdered, injured, destruction, attacking, torn, murders, bomb, killing, abused, punish, committing, abuses, assault, engaging, racist, destroy, harm, weapon, fatally, hacked, trapped, suffering, exhausted, stolen, unacceptable

Continued on next page

Table C.1 – *Continued from previous page*

Moral Foundation	**Category**	**Words List**
Care / Harm	Virtue	compassion, welcomed, friendly, treat, friendship, joy, emotional, rescue, ideals, safe, costly, entertainment, carefully, save, restore, love, capable, respect, saved, improving, create, helping, promoting, wish, ease, relief, help, welcome, value, convinced, growth, safety, expand, care, helps, beautiful, desire, nice, natural, peace
Fairness / Cheating	Vice	rigged, undermining, punished, steal, inability, impose, violating, violated, unfair, injustice, sentenced, violation, struggles, sentences, fake, lies, worries, devastating, threatened, deny, destroying, punishment, low, hide, refuse, delayed, accused, victims, fails, solutions, detained, winners, fraud, threaten, detention, weak, prevented, suspended, pay, illegal
Fairness / Cheating	Virtue	freedoms, fair, honored, wealthy, integrity, recommend, advantage, entitled, swiftly, ethical, grant, protect, motivated, peacefully, accept, awarded, promises, defender, interests, benefit, freed, free, ensure, accepting, dump, legally, honor, opportunities, grants, supports, benefits, wealth, winner, giving, attract, agreed, engagement, enjoy, ensuring, freely
Loyalty / Betrayal	Vice	betrayal, fearful, stealing, hateful, rejection, excuse, hatred, prejudice, havoc, suspected, abandoned, bias, leaked, burden, rejected, disappointed, injury, blame, accusing, aggressive, offenses, attacker, angry, ignored, conspiracy, careful, struggle, warnings, failure, killer, fear, ban, criticism, reject, affected, uncertainty, warning, bitter, hard, wins
Loyalty / Betrayal	Virtue	loyalty, loves, improvement, solidarity, willingness, outstanding, beneficial, supporter, sharing, rich, unified, improved, hero, confident, passionate, committed, devoted, loved, admit, promised, support, questioned, forgotten, praised, successor, promise, strong, dedicated, commitment, peaceful, gain, advanced, join, pride, solve, confidence, playing, definitely, positive, agree
Sanctity / Degradation	Vice	raping, exploit, rape, victimized, distrust, hates, criminals, racism, denying, destructive, evil, scandal, damaged, tolerance, innocent, terrible, bloody, justified, defeating, dirty, aggression, acceptable, disaster, conflicts, enemy, crazy, struggled, heroin, losing, stress, mistake, feared, isolated, escape, fears, controversial, trouble

Continued on next page

Table C.1 – *Continued from previous page*

Moral Foundation	**Category**	**Words List**
Sanctity / Degradation	Virtue	praise, celebrating, virtues, strengthening, respected, grace, dignity, faith, commitments, clean, values, engaged, truth, brave, valuable, healthy, spirit, promote, resolved, vital, protected, strength, hoping, eager, vision, fit, certain, shared, effectively, dream, truly

Table C.1.: *Full List of Words for each Moral Foundation Sentiment pair*