Alexander Ertl

# Contrastive Pre-Training of Transformer Models for Computational Framing Analysis

## MASTER THESIS

for the attainment of the degree of

Master of Science

submitted to

**Graz University of Technology**

Supervisor:
Assoc. Prof. DI Dr. techn. Elisabeth Lex

Advisor:
Dipl.-Ing. Markus Reiter-Haas

Institute of Interactive Systems and Data Science

Graz, August 2023

# Abstract

Framing is the concept of embedding information into a context so that it can be made sense of. More specifically, in language, it is the highlighting of certain elements of consensus reality that motivate a particular perspective. As such, it has seen much attention within social sciences to determine the effect of media on public opinion. However, manual analysis is neither efficient, nor scalable. Hence, the demand for robust and performant computational analysis tools that are able to work with a relative lack of training data, often in multiple languages.

This thesis presents the contrastive learning framework *mCPT* (multilingual contrastive pre-training of transformers) and analyses the properties of the trained embedding space. I demonstrate the effectiveness of multilingual pre-training and the benefit that contrastive regularisation has on the embedding space in a multi-label setting. While multilingual pre-training is an efficient method of increasing the available amount of training data, contrastive learning enables the model to find better decision boundaries by optimising for uniformity i.e., embeddings are spread over the entire embedding space and alignment i.e., embeddings of samples with similar label vectors are positioned close to one-another.

Superior performance is important, yet increased understanding as to the precise nature of frames is equally indispensable. To this end, I contrast *mCPT* with topics and narratives extracted in an unsupervised manner. While topics efficiently capture *what*, they do not tell us *how* something is presented. Nevertheless, some frames are highly co-occurrent with topics and can be considered as more akin to topics. Narratives consisting of simple subject-predicate-object triples on the other hand, lack contextual information and in some cases, context from multiple surrounding sentences may be necessary for a correct interpretation of the narrative in question.

As such, my key findings are that multilingual pre-training, contrastive learning, and contrast sampling jointly increase performance by increasing training data, reg-

ularising the embedding space, and implicitly oversampling. Specifically, contrastive learning acts upon the embedding space by optimising for uniformity and alignment, also in multi-label settings. With respect to topics and narratives, both are very much distinct from conventional frames but provide useful contextual information that can aid the interpretation of frames.

# Zusammenfassung

In meiner Arbeit beschäftige ich mich mit der Analyse von *Framing*, d.h., der Einbettung von Informationen in einen Kontext, der es erlaubt, diese sinngemäß zu erfassen. Je nach Verwendung der Sprache können dabei gewisse Aspekte der Konsensrealität entweder hervorgehoben werden oder in den Hintergrund treten, wodurch gleichzeitig Probleme oder bestimmte Lösungsansätze suggeriert werden. Daher wurde dem Thema in den Sozialwissenschaften viel Aufmerksamkeit gewidmet, um die Wirkung der Medien auf die öffentliche Meinung zu ermitteln. Da aber die manuelle Analyse einerseits ineffizient und daher auch teuer und andererseits nicht skalierbar ist, ist maschinelle *Framing* Analyse von zunehmender Bedeutung. Die Kosten machen es auch problematisch, große Datenmengen für das Training zu erheben, was hohe Ansprüche an die Herangehensweise stellt. Wir benötigen daher ein Modell, welches mit wenigen Daten, die oft sogar mehrsprachiger Natur sind, umgehen kann.

Zu diesem Zweck präsentiere ich die *mCPT* (mutilingual contrastive pre-training of Transformers) Pipeline und analysiere die Eigenschaften des Einbettungsraums. Ich demonstriere die Effektivität von multilingualem *pre-training* und wie sich *contrastive learning* auf den Einbettungsraum in *multi-label settings* auswirkt. Während multilinguales *pre-training* eine einfache Vergrößerung der Datenmenge ermöglicht, optimiert *contrastive learning* die Struktur des Einbettungsraumes dahingehend, dass Einbettungen von Datenpunkten über den gesamten Raum verteilt werden (*uniformity*), wobei jene Einbettungen, die von Datenpunkten von ähnlichen Klassen stammen, nahe beieinander bleiben (*alignment*).

Zusätzlich zur Beschaffenheit des Einbettungsraums betrachte ich auch das Konzept von *Frames* für sich genauer. Dazu differenziere ich sie von Themen und *unsupervised frames*, welche durch *narrative analysis* erstellt wurden. Ich stelle fest, dass Themen sich ganz wesentlich von *Frames* unterscheiden, da sie nicht erfas-

sen, *wie* etwas porträtiert wird, sondern nur beschreiben, *wovon* etwas handelt. Dennoch gibt es manche *Frames*, die mit gewissen Themen fast immer gemeinsam vorkommen. Diese *Frames* könnten als themenhafter beschrieben werden, da sie eine höhere Ähnlichkeit mit ihnen aufweisen. Narrative, die aus einfachen Subjekt-Prädikat-Objekt-Tripeln bestehen, fehlt es dagegen an kontextuellen Informationen, und in einigen Fällen kann der Kontext aus mehreren umgebenden Sätzen für eine korrekte Interpretation der betreffenden Narrative notwendig sein.

Wesentliche Erkenntnisse meiner Arbeit sind, dass mehrsprachiges *pre-training* die Menge der Trainingsdaten erhöht, während *contrastive learning* auch in *multi-label settings* die Struktur des Einbettungsraums auf *uniformity* und *alignment* optimiert. Die ergänzende Verwendung von *contrast sampling* steigert die Leistung, indem implizit *oversampling* betrieben wird. Schlussendlich zeige ich, dass Themen und Narrative sich ganz deutlich von herkömmlichen *Frames* unterscheiden, aber nützliche Kontextinformationen liefern, die die Interpretation von *Frames* unterstützen können.

# Acknowledgements

First and foremost, I would like to thank the entire research team (Markus Reiter-Haas, Kevin Innerebner, and Elisabeth Lex) participating in the SemEval competition for their excellent work and collaboration. A special thanks also goes to my advisor Markus Reiter-Haas for his continual support and guidance.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Today's vast media landscape and its ability to influence and shape people's opinions makes the analysis of exactly how language is used in this context interesting in multiple research fields (Scheufele, 1999). Within studies of communication and journalism, analysis of the framing of arguments is referred to as *framing* (Entman, 1993). Analysing frames may help detect targeted manipulation and score media sources with regard to their level of subjectivity, which is essential for any political system which values openness and transparency of information flow (Boydstun et al., 2014). Furthermore, a better understanding of language use and more specifically framing can improve communication in the public sphere, which is again a centrepiece of functioning democracies (Entman, 2007). Additionally, analysis of frames in public discourse can afford more targeted solutions to particular issues.

Manual framing analysis is both costly and resource intensive. Furthermore, it can only ever examine a small fraction of the textual information generated every day. The ever increased propagation of information makes efficient computational framing analysis a necessity. However, computational framing analysis is a still unsolved challenge. As such, it requires both advancements to the methods of current framing analysis on multiple fronts. Simultaneously, insights into the nature and quality of the extracted frames are necessary for any empirical evaluation (Ali and Hassan, 2022).

## 1.1 Research Questions

The aforementioned requirements lead to the following research questions:

*How can the training procedure of transformers be optimised to improve computational framing analysis methods?* Computational framing analysis relies on highly nuanced natural language processing (NLP) methods capable of extracting *how* information is presented rather than just *what* is presented (Ali and Hassan, 2022). Furthermore, it is a high-dimensional multi-label classification problem i.e., frames are numerous, and they may co-occur which is inherently challenging. I will explore pre-training and contrastive learning as potential sources for performance gains.

*How does contrastive learning transform the embedding space in multi-label settings?* Contrastive learning is known to optimise for uniformity and alignment on the hypersphere given normalised representations (Wang and Isola, 2020) yet the function of contrastive learning in multi-label settings is not entirely clear. Deeper insights into its behaviour may help design more efficient loss functions and training pipelines.

*How do frames differ from topics, and how should they be interpreted?* For real-world application of framing analysis, it is essential to understand what the extracted frames represent. Ali and Hassan (2022) criticise that computational framing analysis can often be reduced to topic classification. How the frames relate to topics extracted from the same data may provide further insight into the nature of framing. For example, exclusive occurrence of the *Health and Safety* frame in topics related to health would indicate a degradation of the frame to a pure topic.

*Do unsupervised framing detection methods agree with the frames produced by the supervised model and/or do they introduce any new frames?* Narrative analysis can be applied as an unsupervised framing extraction method. Can it help corroborate frames, i.e., do the frames extracted using narrative analysis align well with conventional frames?

## 1.2 Contributions

In sum, my contributions are:

**Contrastive Training Framework** The training framework *mCPT* originally presented in my previous work (Ertl et al., 2023) leverages state-of-the-art transformer models conjoined with a contrastive loss function (HeroCon (Zheng et al., 2022)) adopted from the field of computer vision to conduct data-efficient training on small, multi-label datasets. Training is performed in two phases. Provided with multilingual data, it is capable of exploiting mutual information of samples in all languages by pre-training on all datasets *(phase one)* before the final fine-tuning step on the target language *(phase two)*.

**Analysis of Embeddings** This analysis aims to gain insight into the nature of the contrastive loss function and the resulting transformation of the embedding space in multi-label settings. These results may be conductive to further research on contrastive loss functions in natural language processing (NLP) as well as help determine its limits.

**Analysis of Computed Frames** Given a model capable of extracting frames, it is also essential to determine the frames' semantics, i.e., what information the extracted frames provide us with. To this end, I will conduct a qualitative analysis of the extracted frames, as well as contrasting them to topics produced by a topic model to determine their overlapping and distinct characteristics.

**Exploratory Framing Analysis** Unsupervised framing analysis via narrative analysis can corroborate frames extracted using the supervised model and potentially determine new, previously unspecified frames. Large overlaps in the extracted frames would make a case for the validity of the supervised method and the labelled dataset used to train the model. On the other hand, many new frames would suggest that there are many aspects of frames that are still left out by conventional framing definitions and extraction methods.

## 1.3 Structure

The structure of this thesis will closely follow the research questions and contributions. Section 2 will first examine framing as a whole before narrowing down on the subtopic of computational framing analysis. It will then examine contrastive

training approaches in general, as well as give an overview of contrastive training in NLP. Finally, prior work on framing analysis and methods of narrative analysis will be examined. The Methodology Section 3 will provide a detailed overview of the contrastive training framework *mCPT* and go over the steps conducted for further analysis as well as the datasets used therein. Section 4 will mirror the previous section. It will present an ablation study and further empirical results on *mCPT*, the qualitative analysis of frames, and the analysis of frames using narrative analysis. Finally, Section 5 and Section 6 will discuss insights and avenues for future work.

# Chapter 2

# Related Work

This thesis has three major focal points: *framing analysis*, *contrastive learning* and the *transformation of the resulting embedding space*, and *narrative analysis*. The sections in this chapter will be presented in the same order.

## 2.1 Framing Analysis

Framing is the concept of emphasising or highlighting certain aspects of reality so as to encourage a particular world view, problem definition or opinion (Entman, 1993). As such, analysing framing in media can give important insights on the political bias and goals that journalists or media sources have (Tversky and Kahneman, 1985). Until Scheufele (1999) formalised framing, there existed many fractured definitions. Specifically, they constructed a two-dimensional hierarchy of frames constituting the *independent-dependent* and the *audience-media* dimensions. Frames as *independent* variables are examined in the context of how they influence public opinion, whereas frames as *dependent* variables are analysed so as to determine the influence of public opinion on their use. The second dimension describes them as occurring in media versus in the broad public. In the scope of this thesis, I will focus exclusively on frames in media. The categorisation along the first dimension is not as important for this work, owing to the analysis of frames independent of public opinion.

### 2.1.1 Positivist and Post-Positivist Stances

The philosophical positivist stance posits that objective reality can be measured (Jones and McBeth, 2010). With regard to framing, we must therefore be able to quantitatively analyse certain characteristics of frames to create refutable hypotheses and repeatable experiments. In contrast, the post-positivist stance regards the *qualia* of frames as fundamental because it states that "facts" and "reality" are more often than not social constructs created by many individuals (Jones and McBeth, 2010). As such, the main critique of the purely positivist view, is that frames and narratives cannot be defined from outside their contextual embedding owing to their highly subjective nature. Furthermore, the generalisations demanded by empirical methods tend to ignore micro-contexts and as a result often exclude marginalised groups (Jones and McBeth, 2010). Therefore, I draw upon parallels between frames, topics and narratives to better define them in terms of each other.

### 2.1.2 Computational Framing Analysis

The incredible mass of media makes conventional, manual methods of framing analysis impractical. However, the use of computational methods still faces a number of challenges: a lack of datasets for training and validation and often imprecise definitions of framing (Ali and Hassan, 2022). The latter is partly a consequence of the extremely subjective and context-dependent nature of frames. As such, the attempt to objectively define frames is vulnerable to bias (Jones and McBeth, 2010).

Computational framing analysis methods can largely be divided into three broad categories: *supervised*, *unsupervised* approaches and mixtures thereof. Both the supervised and the unsupervised methods face their own unique challenges. While supervised methods have to work with few and small, labelled datasets, unsupervised methods are hard to train as well as validate. In addition, not all datasets use the same definition of frames and the task is hence reduced to multi-label topic classification (Ali and Hassan, 2022).

**Supervised Computational Framing Analysis** Card et al. (2015) annotate articles using the Policy Frames Codebook (Boydstun et al., 2014) with definitions for 15 frames *(political, health and safety, policy prescription and evaluation, legality constitutionality and jurisprudence, capacity and resources, security and defence,*

*crime and punishment, economic, public opinion, external regulation and reputation, fairness and equality, cultural identity, quality of life, morality, and other)* thus creating the Media Frames Corpus (MFC). The MFC consists of annotated news articles from the domains *smoking*, *immigration* and *same-sex marriage*, each article potentially containing multiple frames.

Liu et al. (2019) create the Gun Violence Framing Corpus (GVFC) and employ a transformer (Vaswani et al., 2017) model to achieve state-of-the-art results. Their main contributions are the application of the BERT (Devlin et al., 2019) language model to the domain of framing detection and the curation of the GVFC dataset. Tourni et al. (2021) extend the dataset to include lead images so as to combine visual and textual information for framing detection, further improving performance.

Task three of the *SemEval 2023* challenge (Piskorski et al., 2023) can also be positioned within supervised computational framing analysis as they annotate a new dataset with the frames of the PFC. In the context of the challenge, Ertl et al. (2023) make use of transformer models and contrastive learning objectives to overcome class imbalances and the small number of training samples. Applying transformer models with contrastive learning also addresses Ali and Hassan (2022)'s second open question: how frames can transcend single documents and form relations across documents. On the one hand, learning good representations is a form of data compression and condensing of the data into parameter matrices. Relations between documents are thus implicitly captured. On the other hand, contrastive learning explicitly performs parameter updates based on losses computed from comparisons between multiple documents.

**Unsupervised Computational Framing Analysis** In their survey of computational framing analysis methods, Ali and Hassan (2022) also examine unsupervised approaches. Specifically, they review approaches based on topic modelling as well as clustering. Topic models capture latent dimensions in the data that are equated to frames. Ali and Hassan (2022) point out that the bag-of-words style topic definitions do not align well with the notion of frames. The clustering-based method (Burscher et al., 2016) relies on word-frequencies as features, and thus also reduces frames to context-independent concepts based purely on word-level features.

**Employing Computational Framing Analysis Methods** Computational framing analysis methods have seen use in multiple practical applications. Mulder et al. (2021) operationalise framing to create a recommender system for the diversification of news suggestions. Specifically, they use frames to increase the diversity of perspective rather than the diversity of features such as source, author or topic. Reiter-Haas et al. (2021) analyse framing of political Tweets and correlation of frames with political parties and thus values. Framing is also relevant to the discrimination between conspiracy theories and science-based news (Fong et al., 2021). They find that language used by Twitter conspiracy theorists and their followers tends to view concepts through the frames of religion, power or death.

In this thesis, I draw both upon previous work on supervised computation framing analysis, in particular Ertl et al. (2023), as well as unsupervised methods. While Piskorski et al. (2023) presented a framing analysis challenge in a supervised setting, unsupervised methods are useful in order to verify and examine the trained model.

## 2.2 Contrastive Learning

Contrastive learning is a learning paradigm with the goal of learning meaningful representations of data. The meaningfulness of the representation is task specific, e.g. an embedding of a document that affords sentiment detection will not necessarily allow for good performance on framing detection tasks. Within the context of the task, however, contrastive learning can improve representations by optimising for uniformity and alignment of normalised representations on the hypersphere (Wang and Isola, 2020). That is, representations should be dispersed over the entirety of the hypersphere (uniformity) thus preserving information while minimising distance between positive pairs (alignment). Furthermore, they show that directly optimising for these two metrics in fact outperforms contrastive learning in many cases. What makes contrastive learning particularly interesting is its applicability in supervised as well as unsupervised settings.

**Supervised Contrastive Learning** In supervised learning, positive and negative pairs are determined on the basis of the samples' classes. Samples of the same class form positive pairs, while samples of different classes form negative pairs. Traditional contrastive losses like triplet loss (Weinberger and Saul, 2009) have been superseded

by batch contrastive losses like SupCon (supervised contrastive) loss (Khosla et al., 2020). SupCon loss is computed on the entire batch and has a number of desirable properties, such as generalising to an arbitrary number of positive samples per batch and scaling contrastive power with an increase to the number of negative samples per class. Furthermore, it implicitly performs hard positive/negative mining, i.e., representations of positive pairs that are significantly different influence the change of representations more than representations that are already similar.

SupCon loss is further extended by Zheng et al. (2022) to be applicable in multi-label settings (Heterogeneous Contrastive Loss or HeroCon). A problem that many contrastive losses not intended for this setting face is that of finding positive pairs. For as the number of classes increases, the chance of finding two samples with identical label vectors diminishes exponentially. To overcome this dilemma, both Zheng et al. (2022) and Su et al. (2022) independently propose label-weighted contrastive losses. The general idea remains equivalent to that of vanilla supervised contrastive learning. Representations of positive samples are positioned close to each other, while representations of negative samples are distanced from one another. In this case, samples are considered positive for a class if they are both positive for that class. However, the degree to which representations of positive samples approximate each other depends on the similarity of their label vectors. While Zheng et al. (2022) propose a similarity measure based on the Hamming distance, Su et al. (2022) employ the cosine similarity.

**Unsupervised Contrastive Learning** Contrastive learning in unsupervised settings cannot rely on label information to find positive pairs. As such, methods applied in computer vision often transform the *same* image along a dimension to which the model should be invariant (Gao et al., 2022). Such transformations include cropping, rotations and colour variations. In this case, all samples that are not a transformation of the original sample are considered negative samples. Some transformations may falsely reduce the distance between negative pairs and need to be considered carefully. If *red cars* and *yellow cars* represent two separate categories, then colour transformations are clearly a poor choice. However, applying the correct transformation may not always be so obvious. Xiao et al. (2021) train multiple classifiers on data augmented by all but one transformation, and perform the final classification via a voting procedure. Unsupervised contrastive learning may

also run into another problem: samples that are wrongly labelled as negative are detrimental to performance. To this end, Chuang et al. (2020) propose a debiasing method that reduces the impact of samples wrongly sampled as negatives.

### 2.2.1 Contrastive Learning in NLP

Contrastive learning has recently seen a lot of success within NLP. Su et al. (2022) apply their multi-label contrastive method to a text classification problem by applying it to transformer models. They use a convex combination of a dense model and a KNN classifier as the head, and show that the contrastively optimised embeddings greatly improve KNN and final classification performance.

A recently identified problem with language embeddings is that embeddings occupy a small region of the vector space, i.e., the *anisotropy* problem. Gao et al. (2022) show that contrastive learning effectively combats this problem by optimising for uniformity, which is closely related. Additionally, they learn state-of-the-art sentence embeddings with SimCSE (simple contrastive learning of sentence embeddings) by working in supervised as well as unsupervised settings. For the former, they work with natural language inference datasets and use contradicting sentence pairs as hard negatives and entailment pairs as positives. In the unsupervised case, they use an identical sentence with two distinct dropout masks as the positive pair. In unsupervised SimCSE they note that it is especially important to start training with an already pre-trained model as it does not increase alignment, only uniformity.

Within information retrieval, contrastive learning has been applied to the training of dense retrievers, i.e., dense neural networks (Izacard et al., 2022). The unsupervised training procedure greatly improves performance of dense retrievers, especially on tasks with little or no (few- or zero-shot) labelled training data. Tunstall et al. (2022) further demonstrate the efficacy of contrastive learning in few-shot settings, proposing SETFIT (Sentence Transformer Finetuning). Their approach vastly outperforms standard fine-tuning procedures. Specifically, they implement a two-stage pipeline sequentially contrastively training the sentence embeddings and the classifier head. Ertl et al. (2023) also make use of a multi-stage pipeline in combination with contrastive learning. Our approach will be explained and analysed below in Sections 3 and 4.

## 2.3 Narrative Analysis

Narrative analysis is the analysis of stories present in communication. Since narrative plays an integral role in human cognition and communication, its analysis with respect to the formation of public opinion has seen increased attention (Jones and McBeth, 2010). However, within NLP it is still a relatively unexplored field. According to Jones and McBeth (2010) policy narratives have

i a *setting or context*

ii a *temporal element* i.e., a beginning, middle and end

iii *heroes, villains and victims*

iv a *moral of the story* that motivates action

Reiter-Haas et al. (2023) extract these elements from text via abstract meaning representations (AMRs) (Banarescu et al., 2013).

AMRs are rooted, labelled graphs with that should be both easy for humans to read as well as easy interpretable by machines. Furthermore, sentences with equivalent meaning should translate to identical AMRs. They were proposed as a unified solution to the diverse set of token labelling and dependency identification tasks. Two limitations of AMRs are however that they are heavily biased towards English and that they do not discriminate between real and hypothetical or imagined events (Banarescu et al., 2013).

The subject-predicate-object tuples extracted using narrative analysis can then be analysed to generate previously unspecified insights (Reiter-Haas et al., 2023). More concretely, this can be used to perform unsupervised framing detection. This approach affords a more nuanced perspective on frames that could not be captured by a simple bag-of-words approach, since it focuses on the message of *who did what*. The application of unsupervised framing analysis will compliment *mCPT* and emphasise its performance, or highlight drawbacks and inabilities to detect certain frames.

In summary, I draw upon the HeroCon contrastive loss function (Zheng et al., 2022) and my previous work (Ertl et al., 2023) which presents the competitive framing analysis model *mCPT*. For the analysis of embeddings, I apply insights on uniformity and alignment gained in Wang and Isola (2020). Finally, I contrast frames

produced by *mCPT* with topics created with BERTopic (Grootendorst, 2022) and narratives (Reiter-Haas et al., 2023).

# Chapter 3

# Methodology

This chapter is divided into four broad sections. It will first go into depth on the model *mCPT* (multilingual Contrastive Pre-Training) presented in Ertl et al. (2023), and then detail the approach applied to the analysis of the embedding space and frames. The last section is devoted to the unsupervised detection of frames.

## 3.1 mCPT

*mCPT* was developed in the context of the framing detection sub-task of the third task of the international SemEval challenge (Piskorski et al., 2023). The challenge aims to advance supervised computational framing methods in multiple languages. The datasets consist of articles with as many as 8 frame labels of the 14 PFC frames. Labelled training data was provided in six of the nine test languages, i.e., three languages presented a zero-shot setting requiring transfer learning. Implicitly, the challenge also demands advances to approaches in dealing with limited data sets, multi-label classification and class imbalances.

### 3.1.1 Evaluation Measures

For the evaluation of the supervised framing detection model, I applied the micro-$F_1$ score as specified by Piskorski et al. (2023). The $F_1$ score (Equation 3.1) is defined as a combination of precision and recall denoted as $P$ and $R$ respectively

$$F_1 = 2\frac{P \cdot R}{P + R} \tag{3.1}$$

where precision measures the ratio of samples that were correctly classified as positive i.e., true positives and all samples that were classified as positive i.e., true positives and false positives (Equation 3.2):

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{3.2}$$

and recall measures the ration of samples that were correctly classified as positive and all samples that are actually positive i.e., true positives and false negatives (Equation 3.3):

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3.3}$$

As such, the $F_1$ score strikes a balance between recall and precision which is of special importance when dealing with highly imbalanced datasets. In multi-label settings, the micro-$F_1$ score is calculated globally by summing over the true positives, false positives, and false negatives of all classes.

### 3.1.2 Pipeline

Various parts of the *mCPT* pipeline (Ertl et al., 2023) shown in Figure 3.1 are responses to the distinct requirements imposed by the setting. Multilingual pre-training attempts to overcome the strongly limited number of training samples by exploiting similarities between datasets in the six different languages. Contrastive fine-tuning learns well-performing representations of the data and is aided by contrastive sampling, which implicitly performs oversampling of less frequently occurring classes. Multilingual pre-training comprises the first *phase* of the training procedure. The model is trained on all six labelled datasets and is the foundation for all further models. The second *phase* makes a distinction between few- and zero-shot languages. Models for few-shot languages go through a second contrastive fine-tuning stage which further improves representations, while zero-shot models continue to train on the available data.

The model itself is based on transformer architectures and uses a dense neural network as the classifier head. Specifically, we employ the multilingual sentence transformer model *paraphrase-multilingual- MiniLM-L12-v2* (Reimers and Gurevych, 2019). The phases themselves consist of multiple stages. As illustrated in Figure 3.1 phase one consists of a *head pre-training* and a *contrastive fine-tuning*

stage while phase two comprises *head pre-training*, *contrastive fine-tuning*, and *head post-training* stages in few-shot settings and a *head post-training* stage in zero-shot settings. Head pre- and post-training stages are identical in concept, although their motivations are distinct. In both cases, the head is trained while the transformer body is kept frozen, i.e., the weights are not updated. Since the head is randomly initialised for phase one as well as phase two, directly propagating the initially high gradients through the head into the body would mangle the embeddings of the pre-trained transformer. Once the gradients have diminished somewhat as the head begins to fit the data, we begin the *contrastive fine-tuning* stage. Due to the fact, that the *contrastive fine-tuning* stage updates both the head and the body, the head cannot fully fit the data since the underlying representations keep changing. Therefore, upon completion of the contrastive stage, we again freeze the body and train the head until convergence.

### 3.1.3 Contrastive Fine-Tuning

Supervised contrastive fine-tuning is centred around the idea that latent representations of samples of the same class should be similar, while representations of samples with different classes should differ. This does not translate to multi-label contexts because the number of distinct label vectors grows exponentially with each new class. As such, the likelihood of finding positive pairs, i.e., samples with equal label vectors, approaches zero when the number of classes increases. To avoid this problem, both Zheng et al. (2022) and Wang et al. (2022) propose weighting of losses by similarity measures of their label vectors. Figure 3.1 (top) illustrates how samples with similar label vectors attract each other, while samples with distinct label vectors repel each other.

In *mCPT* (Ertl et al., 2023) we adopt HeroCon loss described in Zheng et al. (2022). Accordingly, the loss function 3.5 comprises a linear combination of two terms: a binary cross entropy term $\mathcal{L}_{BCE}$ which is propagated through the dense classifier and a contrastive term $\mathcal{L}_{CON}$ weighted by an adjustable hyperparameter $\alpha$.

$$\mathcal{L} = \mathcal{L}_{BCE} + \alpha\mathcal{L}_{CON} \tag{3.4}$$

Equation 3.5 describes the contrastive loss term. $C$ denotes the set of classes or, in this case frames, $\mathcal{P}(c)$ is the set of all samples $X$ that are instances of a class i.e.,

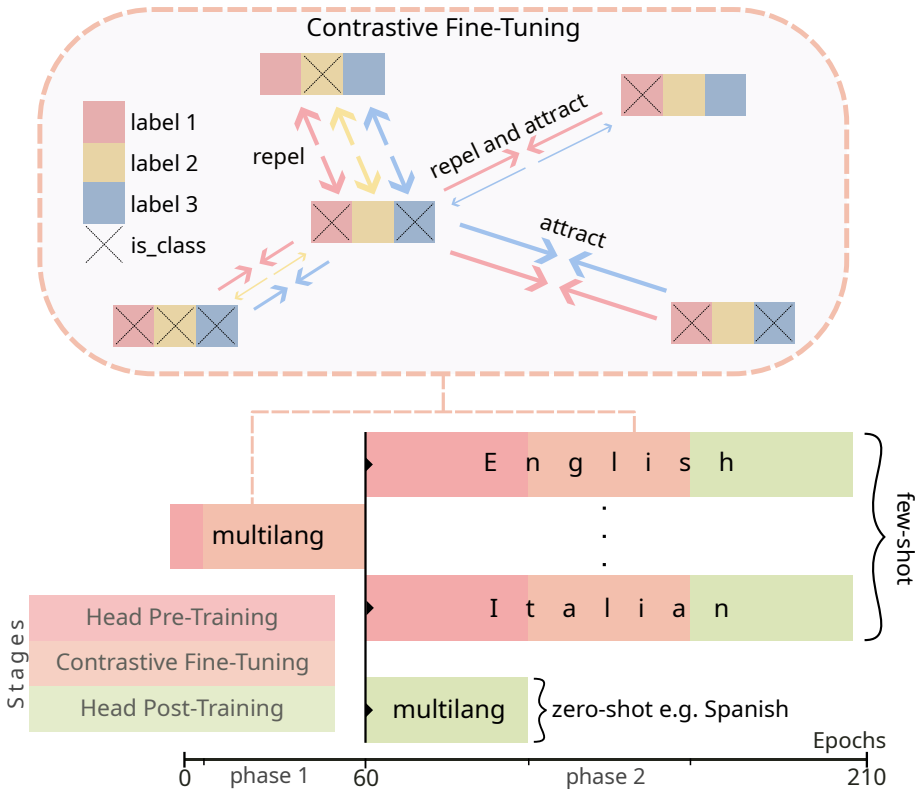



Figure 3.1: A schematic of *mCPT* (from Ertl et al. (2023)). *mCPT* is trained in a two-phase multi-stage procedure: *phase 1* trains the model on all labelled languages, *pre-training* the head with the base transformer model before proceeding to the *contrastive fine-tuning* stage. The resulting transformer body is directly used for zero-shot languages, whereas it is further updated for few-shot languages. *contrastive fine-tuning* applies the contrastive loss function. Embeddings of samples with similar label vectors attract each other, whereas embeddings of samples with different label vectors push each other away.

samples that are of class $c$ whereas $\mathcal{N}(c)$ is the set of samples $X$ that are not of class $c$. $f(\cdot,\cdot)$ is a similarity measure between samples, in this case, the cosine similarity.

$$\mathcal{L}_{CON} = \frac{1}{|C|}\sum_{c\in C} -\mathbf{E}_{X_i,X_j\in\mathcal{P}(c)}\left[\log\frac{\sigma_{ij}f(X_i,X_j)}{\delta_{ij}}\right] \tag{3.5}$$

The denominator $\delta$ is given in Equation 3.6 for readability.

$$\delta_{ij} = \frac{\sigma_{ij}f(X_i,X_j) + \sum_{X_k\in\mathcal{N}(c)}\gamma_{ik}f(X_i,X_k)}{|\mathcal{N}(c)|+1} \tag{3.6}$$

$\sigma$ and $\gamma$ (Equation 3.7) are weights based on the label vectors $Y$ and $Y'$ of the corresponding samples $X$ and $X'$. In particular, $d(\cdot, \cdot)$ denotes the Hamming distance.

$$\sigma_{ij} = 1 - d(Y_i, Y_j)/|C|, \quad \gamma_{ik} = d(Y_i, Y_k) \tag{3.7}$$

It is useful to rewrite the log term of $\mathcal{L}_{CON}$ 3.5 as

$$\log(\sigma_{ij} f(X_i, X_j)) - \log(\delta_{ij}) \tag{3.8}$$

This confirms our intuition that the loss becomes large for batches with many negative pairs with similar representations as well as label vectors, whereas it decreases for batches with many positive pairs that have similar representations and label vectors.

### 3.1.4 Multilingual Pre-Training

Due to the aforementioned class imbalances and small datasets, multilingual pre-training proves to be an effective way of increasing the amount of training data while not compromising on task-specificity. While the label distributions are distinct between the languages, and it is also likely that the language distributions would differ, it is nevertheless plausible that using training data from all languages should result in very generalisable embeddings. In few-shot settings, phase two further fine-tunes the transformer body to adjust for any discrepancies between the global and target-language distributions.

### 3.1.5 Contrast Sampling

Due to the nature of the HeroCon loss function and more generally contrastive training, every batch should ideally have at least one positive pair or at least sample per class. This can not be guaranteed when sampling randomly and in the case of class imbalances becomes highly unlikely when combined with small batch sizes which are often imposed by resource limitations. It therefore makes sense to sample randomly from every class individually when constructing batches. Sampling in this way also implicitly performs oversampling of less frequently occurring classes, since one epoch ends only when every sample has been seen at least once.

## 3.2 Embeddings Analysis

Wang and Isola (2020) show that contrastive learning asymptotically optimises uniformity and alignment. I show that *mCPT* also optimises for these criteria in multi-label settings, building upon work done in Ertl et al. (2023). Furthermore, by applying the loss function to randomly generated samples in low dimensions, I confirm the intuitive understanding of how contrastive learning transforms the embedding space (Ertl et al., 2023). Specifically, the initial embeddings of the random samples stem from a continuous uniform distribution in the interval $[0, 1)$ in three and two dimensions. In the three-dimensional case, the embeddings are plotted using PCA. The three-dimensional label vectors are sampled such that there are 20 samples of a single class, 8 samples of two classes and 8 samples of all three classes. The samples are then iteratively re-positioned by applying gradient descent.

## 3.3 Frame Analysis

As previously discussed, the subjective and contextual nature of frames makes objective analysis impossible. As such, I will apply two distinct approaches to frame analysis. Firstly, I will conduct a qualitative analysis by examining words indicative of frames (*frame keywords*) by examining the coefficients of a linear regression model trained to approximate *mCPT*. This will further understanding of the precise nature of frames annotated for the SemEval challenge. Secondly, I contrast topics and frames. If the frames extracted were found to be equivalent to topics extracted from the same articles, they would be entirely superfluous. To this end, I fit the topic model BERTopic (Grootendorst, 2022) to the articles, as well as applying the framing model fitted on the SemEval dataset. We can then examine the conditional probabilities of the generated topics and frames: $\mathcal{P}(\text{frame}|\text{topic})$ and $\mathcal{P}(\text{topic}|\text{frame})$. A high bidirectional conditional probability would indicate a high degree of similarity between the topic and the frame. By comparing the topics to *frame topics,* i.e., PFC definitions of frames that were made more topic-like (procedure below) we can ultimately see how closely frames and topics relate.

*Frame topics* are defined by removing everything from the original PFC definitions not referring to a concrete physical entity or concept, i.e., everything that is to a degree subjective while maintaining as much of the original wording as possi-

ble. The original frames (Boydstun et al., 2013) and my derivations are listed below. Note that my definitions in some cases use much of the wording found in the original definitions so as to avoid distortion.

1. "*Economic frame: The costs, benefits, or monetary/financial implications of the issue (to an individual, family, community, or to the economy as a whole).*"

   Topics related to monetary/financial issues.

2. "*Capacity and resources frames: The lack of or availability of physical, geographical, spatial, human, and financial resources, or the capacity of existing systems and resources to implement or carry out policy goals.*"

   Topics related to physical, . . . , and financial resources.

3. "*Morality frames: Any perspective—or policy objective or action (including proposed action)that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other sense of ethics or social responsibility.*"

   Topics related to religion and ethics.

4. "*Fairness and equality frames: Equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups. Also the balance between the rights or interests of one individual or group compared to another individual or group.*"

   Topics related to fairness and equality.

5. "*Constitutionality and jurisprudence frames: The constraints imposed on or freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights and other amendments, or judicial interpretation. This deals specifically with the authority of government to regulate, and the authority of individuals/corporations to act independently of government.*"

   Topics related to the Constitution, Bill of Rights and other amendments.

6. "*Policy prescription and evaluation: Particular policies proposed for addressing an identified problem, and figuring out if certain policies will work, or if existing policies are effective.*"

   Topics related to the evaluation of implemented or not yet implemented policies.

7. "*Law and order, crime and justice frames: Specific policies in practice and their enforcement, incentives, and implications. Includes stories about enforcement and interpretation of laws by individuals and law enforcement, breaking laws, loopholes, fines, sentencing and punishment. Increases or reductions in crime.*"

   Frame definition serves as topic.

8. "*Security and defense frames: Security, threats to security, and protection of one's person, family, in-group, nation, etc. Generally an action or a call to action that can be taken to protect the welfare of a person, group, nation sometimes from a not yet manifested threat.*"

   Topics related to the concepts of the first sentence of the frame definition.

9. "*Health and safety frames: Healthcare access and effectiveness, illness, disease, sanitation, obesity, mental health effects, prevention of or perpetuation of gun violence, infrastructure and building safety.*"

   Topics related to concepts in definition.

10. "*Quality of life frames: The effects of a policy on individuals' wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc.*"

    Topics related to individual wealth, mobility, ...

11. "*Cultural identity frames: The social norms, trends, values and customs constituting culture(s), as they relate to a specific policy issue.*"

    Topics related to social norms, ... and customs constituting culture(s).

12. "*Public opinion frames: References to general social attitudes, polling and demographic information, as well as implied or actual consequences of diverging from or "getting ahead of" public opinion or polls.*"

    Topics related to public opinion, social attitudes, and demographic information.

13. "*Political frames: Any political considerations surrounding an issue. Issue actions or efforts or stances that are political, such as partisan filibusters, lobbyist involvement, bipartisan efforts, deal-making and vote trading, appealing*

*to one's base, mentions of political maneuvering. Explicit statements that a policy issue is good or bad for a particular political party.*"

Topics related to the concepts of the first sentence of the frame definition.

14. "*External regulation and reputation frames: The United States' external relations with another nation; the external relations of one state with another; or relations between groups. This includes trade agreements and outcomes, comparisons of policy outcomes or desired policy outcomes.*"

    Frame definition serves as topic.

Finally, I calculate topic coherence measures on 14 topics created by BERTopic and the 30 most significant words for classifying frames extracted by the linear model. High coherence measures of these *frame keywords* would demonstrate an inherent similarity of frames according to the definitions of the PFC and topics. The coherence measures are calculated using the GenSim coherence model based on the best measure $C_V$ determined in Röder et al. (2015). The general idea is to measure the coherence of a topic, i.e., a set of words, by evaluating how strongly subsets "hang" together. Like all coherence measures examined in Röder et al. (2015) $C_V$ comprises a segmentation method, a confirmation measure used to score the agreement of a pair of words or sets, and a probability measure. The segmentation method partitions the topics. These partitions are then evaluated by the confirmation measure, in this case *normalised pointwise mutual information*. Pointwise mutual information is in turn based on the ratio of the probability of two words occurring together versus their probabilities of occurring altogether. Finally, the coherence of the entire topic is calculated by aggregating all partition confirmation measures via the arithmetic mean.

## 3.4 Unsupervised Frame Detection

Similarly to Reiter-Haas et al. (2023), I perform unsupervised frame detection via narrative analysis and extraction of subject-predicate-object tuples. To find relevant tuples, i.e., frames, I additionally apply filtering operations based on the number of global and local (per document) occurrences as well as sentiment.

My approach is based on the following assumptions:

i Strong frames, especially ones that were explicitly crafted to hammer home a message, will occur multiple times within a single document

ii The dataset is not homogeneous i.e., relevant frames will only occur in a subsection of the documents

iii Frames will often employ high-valence words that convey strong emotion

**Narrative Analysis** Narrative analysis is performed using AMRlib[1] for generation of the abstract meaning representations and the Penman library (Goodman, 2020) for mining of the narratives themselves. As in Reiter-Haas et al. (2023), I extract narrative elements i.e., the character (ARG0 / subject), the plot (predicate), and the objects if they exist (ARG1) by traversal of the graph's edges. The predicate relates the subject and the first object and is therefore the parent node. However, the first object may have further child nodes, without which the frame might become meaningless. Simultaneously, this could make frames too specific. To address this issue, I consider both base and extended tuples.

**Sentiment** The valence of frames is determined using the Vader sentiment detection library (Hutto and Gilbert, 2014). As per *assumption iii*, frames are discarded if they have neutral compound score. The score of a narrative is a weighted sum of valence scores of all words occurring in the narrative and normalised to be between -1 and 1. Valence scores are determined using a lookup table, i.e., a dictionary. The weights for the compound computation stem from a rule base specifying adjustments based on context.

**Interpretation** Similarly to our prior analysis of topics and frames, it is informative to contrast narratives and the frames generated with *mCPT*, this time by examining mutual information. Mutual information (Equation 3.9) is a measure for how much information about one random variable (RV) is gained by the knowledge of another. Note that it does not tell us whether the realisation of one RV makes the realisation of another more or less likely.

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \left( \frac{P(x,y)}{P(x)P(y)} \right) \tag{3.9}$$

[1] `https://github.com/bjascob/amrlib`

Equation 3.9 is a double sum over the two-dimensional matrix containing all instances of $X$ and $Y$. Examining the absolute values of the matrix is difficult as the values have no upper bound. However, relatively high mutual information (as compared to other values in the matrix) is an indication of the relatedness of a narrative and the corresponding frame.

# Chapter 4

# Experiments and Results

All evaluations, aside from the ablation study, used *mCPT* in an identical configuration. The body is the *paraphrase-multilingual- MiniLM-L12-v2* transformer model and the head a three-layer dense neural network with a single hidden layer of size $256x256$ with ReLU activations. Model fine-tuning was performed on Kaggle (`www.kaggle.com`) on P100 graphic cards, which also determined the batch size of 26. The maximum batch size in turn also constrained contrast sampling. Given the batch size and the number of classes (14) we can only require one positive sample per batch, thus only guaranteeing at least one negative pair per class (instead of one negative and one positive pair). This is however already sufficient to increase performance and consistency. In all cases, the model was trained using identical hyperparameters (Table 4.1) which included separate learning rates $\eta$ as well as decay rates $\gamma$ for the head and the body. The entire code is available publicly at `https://github.com/lambdasonly/mCPT`.

Table 4.1: Model hyperparameters. The head and body were trained using different learning rates $\eta$ and decay rates $\gamma$. $\alpha$ weights the contrastive regularisation term of the loss function (Equation 3.4).

| head-$\eta$ | head-$\gamma$ | body-$\eta$ | body-$\gamma$ | $\alpha$ |
|---|---|---|---|---|
| $1 \times 10^{-3}$ | 0.99 | $2 \times 10^{-5}$ | 0.98 | 0.01 |

## 4.1 Datasets

*mCPT* was evaluated purely on the dataset created by Piskorski et al. (2023). There were six training and corresponding development datasets in the languages English, Italian, Russian, French, German and Polish with between 433 and 132, and 83 and 45 articles (details in Table 4.2). Within the context of the challenge, all models were evaluated on nine hidden test datasets which have not yet been released. The test datasets are in the six aforementioned languages as well as three zero-shot languages: Spanish, Georgian, and Greek. As previously stated, the dataset is

Table 4.2: Number of training/development/test samples in the SemEval (Piskorski et al., 2023) dataset.

| | English | Italian | French | Polish | Russian | German |
|---|---|---|---|---|---|---|
| train | **433** | 227 | 158 | 145 | 143 | 132 |
| dev | **83** | 76 | 53 | 49 | 48 | 45 |
| test | 54 | 61 | 50 | 47 | **72** | 50 |

highly imbalanced. Figure 4.1 illustrates the extent of the class imbalances as well as shifts in distribution on three of the six train and development datasets.

Frame analyses were conducted on a random sample of the LOCO (Miani et al., 2021) conspiracy news dataset. 10 000 of the 96 743 articles were sampled with `sklearn.utils.random.sample_without_replacement` (Buitinck et al., 2013) setting a random seed of 42. The dataset comprises mainstream as well as conspiracy documents from over 150 websites.

## 4.2 Ablation Study

The ablation study (results in Table 4.3) was conducted on the SemEval development sets. It demonstrates the effectiveness of all components of the *mCPT* training procedure. *mCPT* with the contrast sampling extension shows superior performance on four of the six test languages, while only coming in second by a small margin on the remaining two. Figure 4.2 illustrates that contrast sampling, while not decreasing variance, slightly boosts performance. The micro-$F_1$ scores of ten training runs, conducted on the English training set and evaluated on the English development set using different seeds, show that contrast sampling is on average able to increase

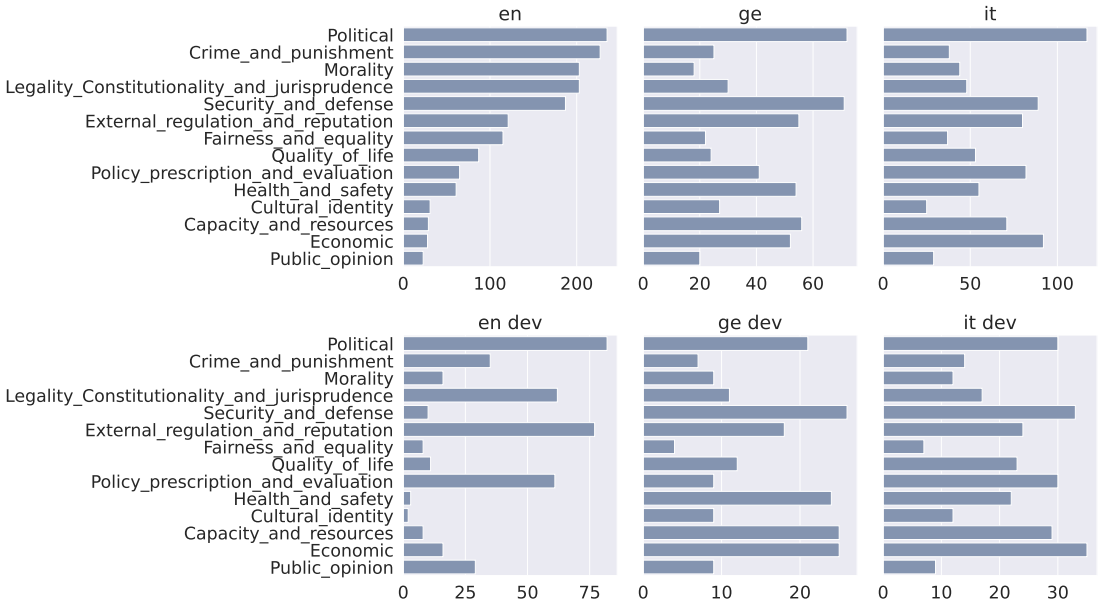



Figure 4.1: Frame counts on the English, German, and Italian train and development datasets. Class imbalances and shifts in train and development distributions require robust learning methods.

performance by 0.03 points. Removing pre-training (- PT) and contrastive learning (- $\mathcal{L}_{CON}$) further degrades performance. These drops are likely due to the strong size reduction of the training dataset on the one hand, and the fact that contrastive learning may be able to make better use of the data on small datasets than conventional training methods, on the other (Tunstall et al., 2022). The inconclusive results on end-to-end training (E2E) are perhaps similarly related to the relatively small dataset and the under-performance of standard fine-tuning methods in such cases (Tunstall et al., 2022).

## 4.3 Analysis of Embeddings

In this section I analyse the structure of the embeddings of the initial transformer model, the pre-trained model and the fully trained language-specific model (trained on the English SemEval dataset). Results confirm that multi-label contrastive learning also optimises for uniformity and alignment. Furthermore, a toy example demon-

Table 4.3: The ablation study performed in Ertl et al. (2023) shows that all components of the training pipeline are required to achieve best results. $\overline{\Delta}$ denotes the average performance loss upon removing a component. Performance losses by removing contrastive sampling (- CS), pre-training (- PT), and contrastive learning (- $\mathcal{L}_{CON}$) are especially noteworthy. However, end-to-end training (E2E) on its own does not always result in performance gains. This emphasises the necessity of contrastive learning in addition to end-to-end training.

| Model | en | it | ru | fr | ge | po | $\overline{\Delta}$ |
|---|---|---|---|---|---|---|---|
| *mCPT*+CS | **.688** | **.590** | .519 | **.575** | **.591** | .638 | |
| - CS | .682 | .585 | **.520** | .570 | .561 | .636 | -.008 |
| - PT | .681 | .545 | .475 | .563 | .583 | .616 | -.015 |
| - $\mathcal{L}_{CON}$ | .657 | .521 | .436 | .524 | .570 | **.645** | -.018 |
| - E2E | .629 | .519 | .500 | .535 | .586 | .633 | .008 |

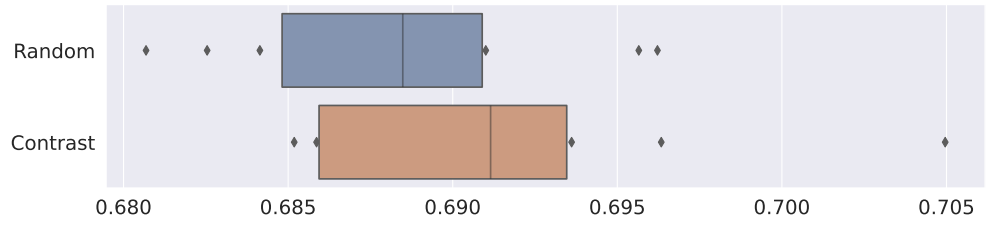
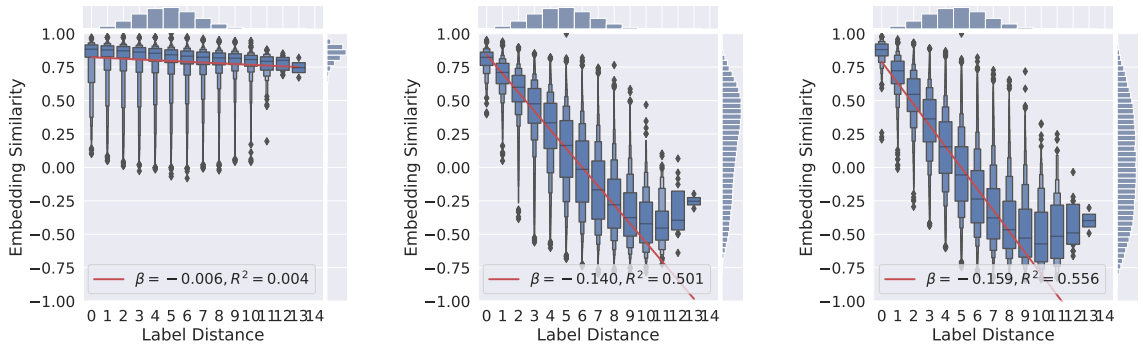



Figure 4.2: Micro-$F_1$ scores for contrastive and random sampling methods applied to the English development dataset using different seeds. On average, contrast sampling boosts performance by 0.03 points.

strates operation of the contrastive loss function in a low-dimensional space that affords visualisation.

Figure 4.3 plots cosine similarities of embeddings against the Hamming distances of the label vectors. It illustrates increased similarity of embeddings among samples with low label vector Hamming distances and decreased similarities of embeddings with high label vector Hamming distances after contrastive learning. This is conceptually supported by Wang and Isola (2020) as higher embedding similarities of similar samples are related to higher alignment. It is noteworthy, that prior to any training, almost all cosine similarities are positive, i.e., similar to a degree. In contrast, cosine similarities of contrastively trained embeddings may also be negative, i.e., dissimilar. While the figures indicate that the multilingual pre-training step

leads to the largest increase in embedding separation, the regression coefficient $\beta$ shows that target-language-specific training further increases separation (given the further decrease from $-0.14$ to $-0.16$).




(a) Untrained (b) Pre-trained (c) Trained

Figure 4.3: Label vector Hamming distances plotted against the cosine similarities of the embeddings (from Ertl et al. (2023)). Contrastive training decreases the similarity between samples with high Hamming distances, while preserving the similarities between samples with similar label vectors.

Wang and Isola (2020) show that contrastive learning asymptotically optimises alignment as well as uniformity, yet they only apply their analysis to normalised embeddings in non-multi-label settings. While Figure 4.3 demonstrates increased alignment, Figure 4.4 further supports the claim that *mCPT* generates more uniformly distributed embeddings. The histograms represent distributions of the co-variances of the normalised sample-embeddings. The histogram on the left is notably shifted to the right because of a high number of positive co-variances. This indicates that most samples are located in a similar region of the hypersphere. Contrastive training shifts the distribution towards the centre and decreases mean co-variance due to a more uniform spread of sample-embeddings over the hypersphere.

A toy example (Figure 4.5) shows exactly how the contrastive loss function transforms the embedding space. Samples with identical label vectors are oriented along axes drawn from the origin. Intuitively, this makes sense because the loss function decreases the cosine similarity between samples with similar label vectors. Furthermore, embeddings are *close* to each other, given *similar* label vectors, e.g. samples of only class *yellow* are closer to samples of class *yellow* and *red* than to samples of class *red* and *blue*. This means that every unique label vector is represented by

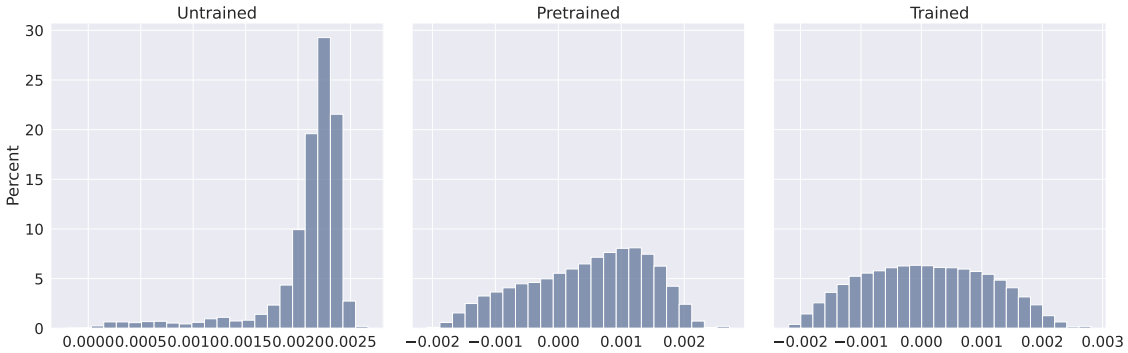



Figure 4.4: Histograms of the co-variances of the normalised sample-embeddings. High average co-variances are the result of the samples being located in a similar section of the hypersphere, i.e. most signs are equal, so the samples are located in a small cone. Contrastive training decorrelates the sample embeddings, indicated by a more uniform distribution.

an axis within the embedding space. If we want our embedding space to be structured such that two axes are closer to each other if the label vectors they represent have a lower Hamming distance, it is a requirement that the dimensionality of the embedding space be at least as high as the number of classes. We already run into this problem in the toy example: *red-yellow*, *red-blue* and, *yellow-blue* should all be equally close to *red-yellow-blue* while all being equidistant. Simply superposing them is no solution, as *blue* should be closer to *red-blue* than *red-yellow*. As such, some axes have merged. Adding a dimension to the embedding space solves this problem by allowing axes to have distinct depth values. Figure 4.6 illustrates a 3d example projected into two dimensions using PCA. Even though it is a projection, it shows clearer separation than the 2d example. The only axes with overlap (*red-yellow-blue* and *red-blue*) should be considered to have distinct depth values, i.e., they do not overlap in three dimensions.

**Euclidean Contrastive Loss** Not all distance measures work equally well in the context of contrast learning. The contrast loss function in Equation 4.1 (Su et al., 2022) is based on the Euclidean distance $d$ between embeddings $z$ weighted

(a) Untrained (b) Trained

Figure 4.5: Application of the contrastive loss function to a two-dimensional toy example. The colours represent different classes, while the saturation of the colour indicates whether the sample is positive for the respective class or not. Contrastive training aligns samples with identical label vectors along axes drawn from the origin. In this particular case, the dimensionality of the embedding-space is not sufficient to capture the structure in the label-space (see text).

by normalised dot products of their label vectors $y$ (Equation 4.2).

$$\mathcal{L}_2^{ij} = -\beta_{ij} \log \frac{\exp(-d(z_i, z_j)/\tau')}{\sum_{k \in g(i)} \exp(-d(zi, z_k)/\tau')} \tag{4.1}$$

$\tau'$ denotes the learning temperature and $g(i)$ returns all indices of a batch not equal to $i$.

$$\beta_{ij} = \frac{y_i^\top y_j}{\sum_{k \in g(i)} y_i^\top y_k} \tag{4.2}$$

Use of the dot product rather than the Hamming distance means that samples that are both negative for the same class are not considered more similar to each other and pairs with one negative and one positive sample are not considered as more dissimilar i.e., $\beta_{ij}$ is large *iff* $y_i$ and $y_j$ are both positive for the same classes. As such, negative pairs are not explicitly pushed away from each other because they are not even considered. However, due to the denominator of Equation 4.1 all samples are implicitly pushed away from each other i.e., the denominator ensures that the embedding space does not implode.

The Euclidean-based distance results in a very distinct transformation of the embedding space. Rather than aligning samples with identical label vectors along axes drawn from the origin as in Figure 4.6 samples with similar label vectors are

Figure 4.6: The untrained 3d embeddings are similarly chaotic as the untrained 2d embeddings, however, in the trained case, all unique label vectors are clearly separated. In the 2d PCA projection only *red-yellow-blue* and *red-yellow* overlap slightly. In 3d, the 1-class, 2-class and 3-class samples all have distinct depth values.

clustered (Figure 4.7). Samples with label vectors of high magnitude are located towards the origin, while samples with label vectors of low magnitude are located at the periphery in order to maximise distance between each other.

This loss function did not perform as well as Equation 3.5 which was instead based on the cosine similarity between embeddings. Euclidean distance is known to degrade in high dimensions, it is therefore possible that the sparsity of the resulting embedding space does not lend itself as well to classifiers.

**Usability of Frame Embeddings on NLP Tasks** The nature of sentence embeddings that are averages of the individual token embeddings means that models trained for sentence classification will not perform well on tasks that require token level information. Individual token embeddings are in fact transformed in such a way that, after training, not averaging the token embeddings would not lead to a dramatic decrease in performance i.e., directly applying the classification head to individual tokens would in most cases correctly predict the frames occurring in the

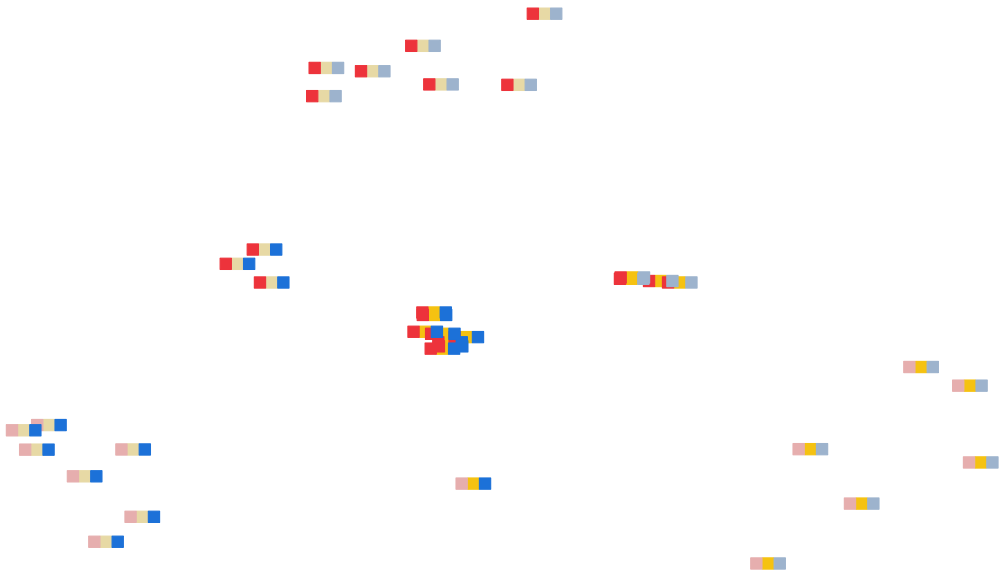


Figure 4.7: 3D embeddings projected using PCA after training using a Euclidean distance-based contrastive loss function (Equation 4.1). Samples with high-magnitude label vectors are central, while samples with low-magnitude label vectors are located at the periphery. As such, distances between samples of common classes are minimised.

entire article.

## 4.4 Frame Analysis

Table 4.4 lists the 15 terms that are most significant for predicting the corresponding frame, i.e., the first 15 *frame keywords* as well as the linear model's micro-$F_1$ score on the fitted data. For brevity, only the 8 highest $F_1$ scores are listed. The corresponding frame counts are illustrated in Figure 4.8. Specifically, the linear model was trained on the output of a 1-3-gram (1, 2, and 3-grams were considered) *TF-IDF* transformation. Terms with a document frequency lower than 0.001 were discarded, i.e., terms must occur in at least 10 documents. Furthermore, English stop words were removed. Of the listed frames all have micro-$F_1$ scores above 0.6 indicating that the model is able to learn a reasonable approximation of *mCPT* while also showing that a bag-of-words approach is not in itself sufficient. Note that the scores

were calculated on the training set of the linear model.

Table 4.4: The eight best-predicted frames (F1 above 0.6), the 15 most significant predicting words and the corresponding prediction accuracy. While *frame keywords* are somewhat related, they do not form coherent topics.

| Frame | Words | Micro-$F_1$ |
|---|---|---|
| Health and safety | levin, mass genocide, wage, infect human, osteoporosis, converging, somatic, watch tv, guitarist, 890, kidney cells, debra, cares act, people hiv aids, citizens catastrophic | 0.85 |
| Quality of life | parade, dreamy, mass genocide, 24 hours, myth, ahmed, iran nuclear, february 2004, wage, miami, detects, levin, nearly 000, acutely aware, warned world | 0.85 |
| Political | decrees, following statement, far flung, haley, parents child, basket, october 19, lessons, poster, fame fortune, economic advisers, iran nuclear, music festival, pascal, afford send | 0.78 |
| Crime and punishment | foot long, shock, vulnerability, 25 30, straight line, success story, sifting, federal law, strong, previous record, services committee, help lower, live virus, 891, madness | 0.71 |
| Security and defence | meets eye, study funded, search missing, trump win, macron, children health, swirled, just finished, technocracy, marian, american space, governs, harbor, extremes, economic crashes | 0.68 |
| External regulation and reputation | iraqi leader saddam, referral, modified mosquitoes, berlin, pascal, department public, october 19, study funded, malaysian, squeeze, scientific literature, stressors, blindness, senior executive, united states long | 0.68 |
| Capacity and resources | ga, 244, percentage, used treat, lifts, tenuous, blindness, 1956, does people, detects, dramatic increase, said knew, internal organs, student loan, 1812 | 0.66 |
| Morality | tower world trade, perpetrated, caused novel, ecuador, corrosive, success story, denials, los angeles police, backers, proportions, strong, spend billion, intercept, nominate, research suggests | 0.63 |

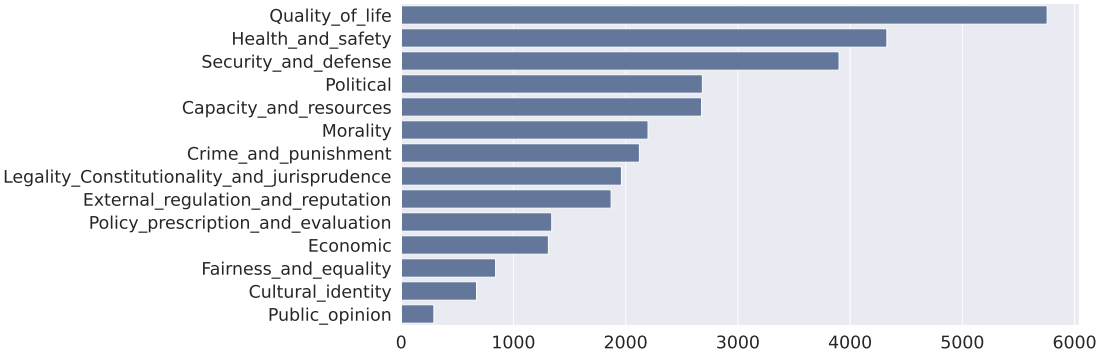



Figure 4.8: The number of frames detected in the random sample of the LOCO dataset using *mCPT*. The chart is as imbalanced as the training dataset, and there are correlations between the samples occurring infrequently in the train and the LOCO dataset. It is not clear whether this is the result of these frames generally occurring less frequently or because the model is less capable of picking up on these frames owing to the lack of training data.

Words are difficult to interpret without context, yet "coloured" terms such as *mass genocide*, *citizens catastrophic*, *acutely aware*, or *far flung* point towards strong framing. Here, it is useful to examine co-occurrences with conventional topics extracted using a topic model. Specifically, the topic model employed is BERTopic with K-means clustering ($k = 30$) to force assignment to clusters rather than marking samples as outliers as would be the case with HDBSCAN (McInnes et al., 2017). Indeed, Figure 4.9 suggests that these terms are used to exaggerate and paint reality in a particular way which is much in line with Entman (1993)'s definition of framing. Since the figure plots conditional probabilities of topics given the frames, the columns are normalised to sum to one and the fields can be interpreted as percentages. For example, the *health and safety* frame mainly co-occurs with topics related to climate change, COVID, or other medical issues and not something that would warrant the use of the phrase *mass genocide*. Closer examination of the use of the term confirms this intuition (own italics):

> With the global warming controversy, Al Gore and his higher up Cabal minions are using global warming in the political realm to push their UN Agenda 21 objectives of total control, ... and ultimately *mass genocide*. (Miani et al. (2021))

> The significance of this US Supreme Court reasoning in their ROE v. WADE ruling legalizing abortion, this report explains, is their having "left the door open" to the outlawing of this horrific *mass genocide* practice ... (Miani et al. (2021))

> Western governments and their intelligence services actively campaign against the information found in these reports so as not to alarm their *citizens* about the many *catastrophic* Earth changes and events to come, a stance that the Sisters of Sorcha Faal strongly disagrees with in believing that it is every human beings right to know the truth. (Miani et al. (2021))

> While the president riffed on the idea in a joking tone, his speech at Prince George's Community College revealed a very serious undercurrent ... The president has few tools to check the rising cost of gasoline in the short term, and his advisors are *acutely aware* of the effect this could have on voters. (Miani et al. (2021))

In the first three cases, the phrase serves to exaggerate or put emphasis on the graveness of the situation while in the third it underlines that officials are taking the problem seriously.

On the other hand, some terms such as *Levin*, *guitarist* or *february 2004* just refer to entities, concepts, or time periods. While it is of course possible to use entities to frame, for example by referring to the views of a political entity, *guitarist* can safely be said to be unrelated to the *health and safety* frame. It is more likely, that the linear model picked up on coincidental patterns such as here:

> ... pioneering rock *guitarist* whose sharp, graceful style helped Elvis Presley shape his revolutionary sound and inspired a generation of musicians ... died Tuesday. He was 84.
>
> Moore died at his home in Nashville, said biographer and friend James L. Dickerson, who confirmed the death .... (Miani et al. (2021))

This limits what we can gain from the linear model. It is however a useful first indicator that *mCPT* frames do not entirely overlap with topics and that the model is applicable to new, previously unseen datasets.

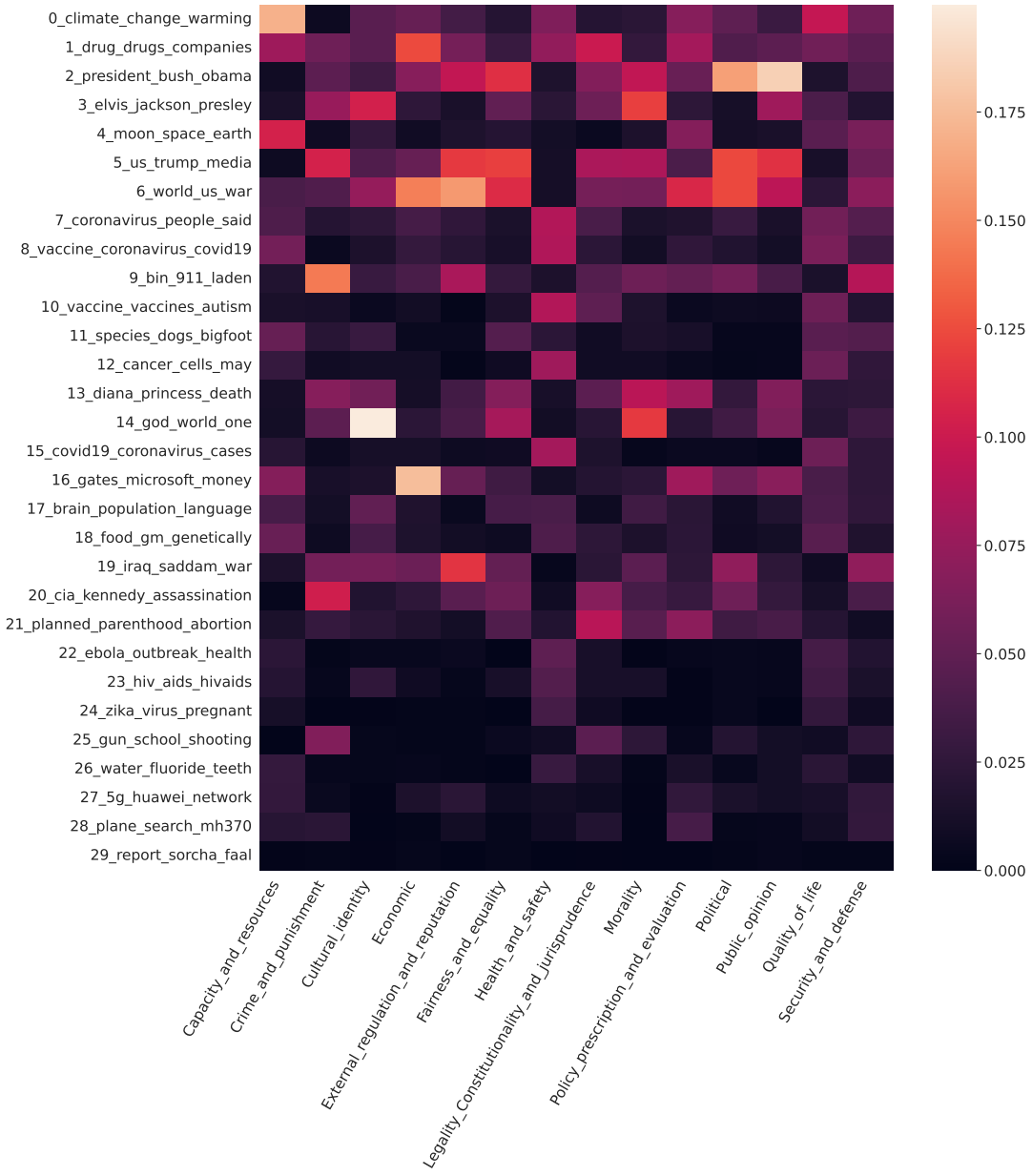



Figure 4.9: The conditional probability of a topic occurring given the frame $\mathcal{P}(\text{topic}|\text{frame})$. The fields of the matrix can be interpreted as percentages, e.g. the documents of topic *14_god_world_one* make up around 20% of the documents with the *cultural identity* frame.

**Contrasting Frames and Topics** Examining the rows of Figure 4.10 which depicts the conditional probabilities of frames given the topics, we can see that almost

all topics utilise multiple frames even though most of them have a predominant frame. This indicates that there is no straightforward mapping from topics to frames. In fact, when comparing topics to the *frame topics* of Section 3.3 some mappings would not be found. For example, the word set corresponding to the topic *0_climate_change_warming* is *climate, change, warming, global, energy, emissions, carbon, temperature, weather, water* and the utilised frames are *quality of life, capacity and resources, health and safety, and security and defence*. Given our prior knowledge of the topic, it is fairly easy to relate to all four frames. However, relating the word set to the *frame topic* definitions is a bit of a stretch. We can associate the words *energy* and *water* with the frame topic *quality of life* via the reference to "effects of a policy on individuals' ... access to resources", likewise for *capacity and resources*. However, there is nothing in the word set that directly allows us to draw parallels to the *health and safety* frame topic. Similarly, the word set for *1_drug_drugs_companies* is *drug, drugs, companies, pharmaceutical, industry, health, prices, company, pharma, said* and documents belonging to the topic utilise almost all frames. However, manually mapping topics to frame topics would not capture all actually found frames i.e., the *legality, constitutionality and jurisprudence* and *security and defence* frame topics are only loosely, if at all, related to the list of words that represents the topic. Another notable example is the topic *21_planned_parenthood_abortion*: *planned, parenthood, abortion, women, abortions, said, wisconsin, health, state, clinics*. The predominant frames are *legality constitutionality and jurisprudence*, *morality*, *policy prescription and evaluation*, and *quality of life*, although one would mainly associate it with the *health and safety*, and *quality of life* frame topics given the frame topic definitions. By applying our knowledge of these topics, we can easily see how a certain *framing* of planned parenthood and abortion might lean heavily on moral arguments. In fact, that is precisely what we set out to determine: whether frames are able to capture information other than a list of words such as a perspective. This shows that topics alone do not encompass frames.

**Average Frame Counts** Having obtained the conditional probabilities $\mathcal{P}(\text{frame}|\text{topic})$, we can sum over the rows of the matrix (Figure 4.10). Since the elements of every row are percentages of how many documents of the topic use a certain frame, the sum represents an average of the number of frames used per

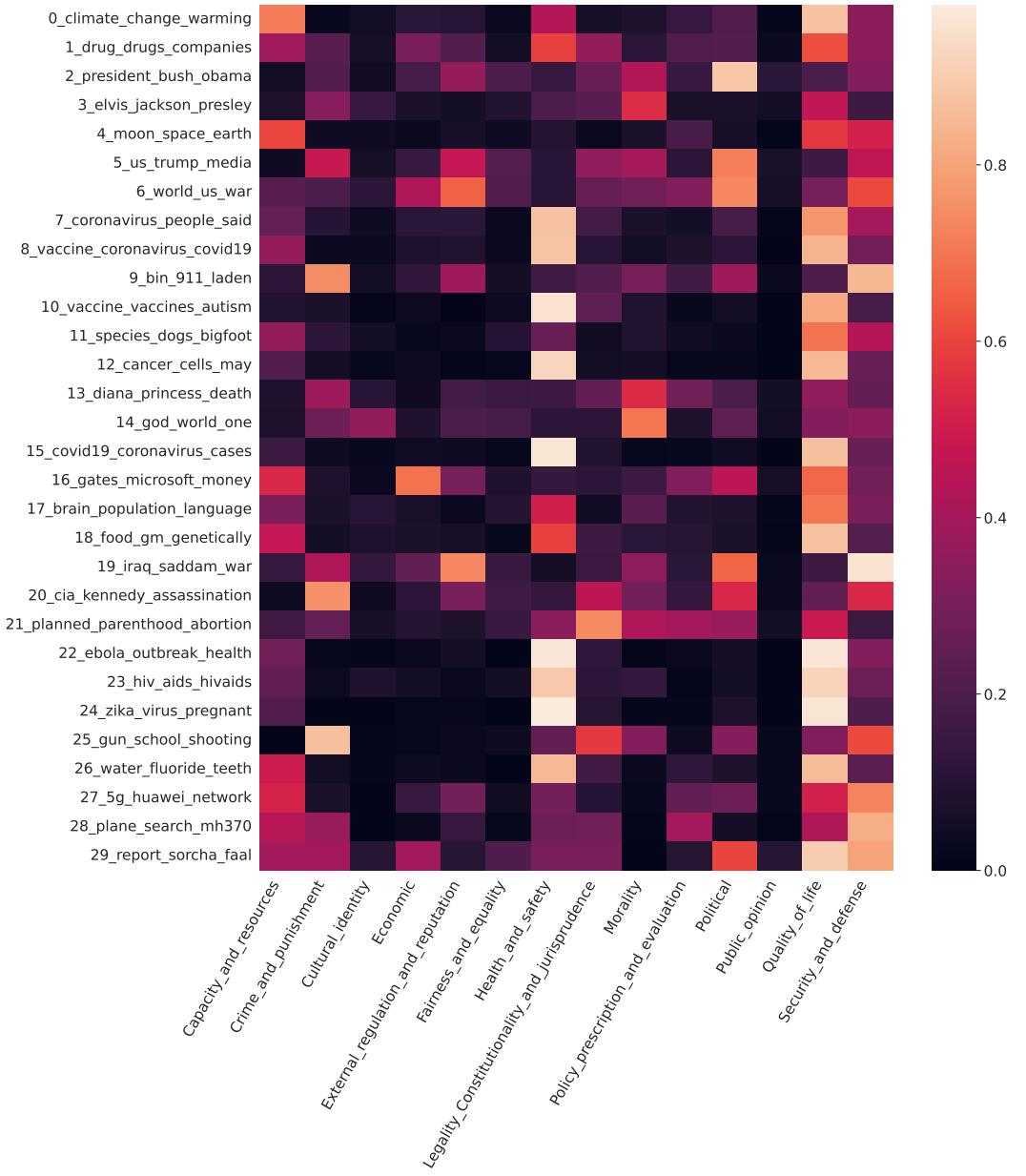



Figure 4.10: Conditional probabilities of frames given topics $\mathcal{P}(\text{frame}|\text{topic})$. This normalisation is useful for examining the relative importance of frames within topics. The presence of some topics, especially those that are health-related, make certain frames almost certain.

document of a particular topic (Figure 4.11). Most of the topics making high use of framing are either related to war, terrorism, or conspiracy theories (Sor-

cha Faal is a known pseudonym for a conspiracy theory blogger under the domain `www.whatdoesitmean.com`). This points towards the emotional nature of frames as all of these topics are highly laden with sentiment. In contrast, the topics using the least frames are, while still related to conspiracy theories (11: *species, dogs, bigfoot...*, 4: *moon, space, earth, alien, apollo, nasa, lunar, landing, life, would*), much less heated.

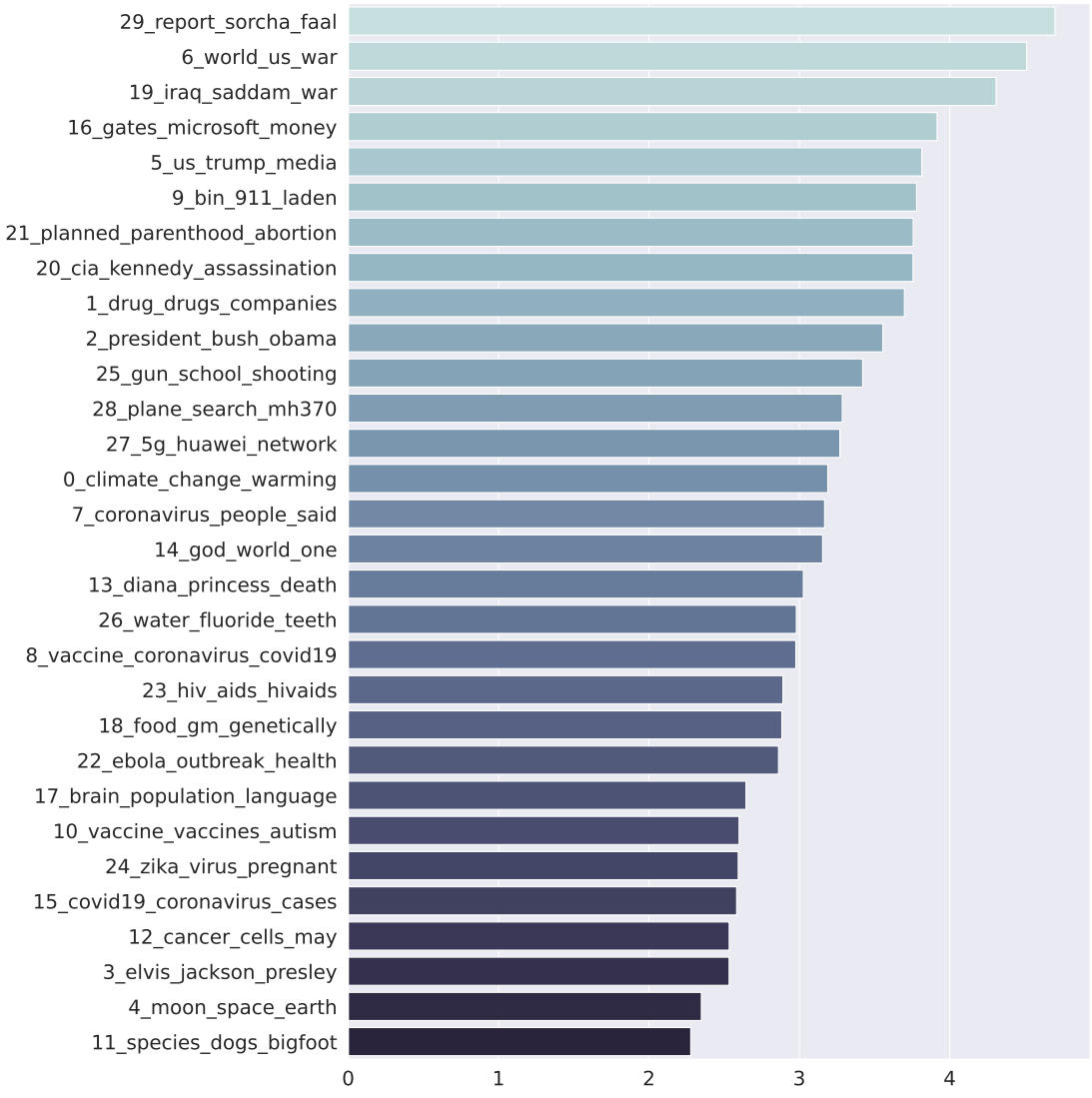



Figure 4.11: Average number of frames per topic. Highly emotional topics revolving around war, terrorism or conspiracy theories tend to use more frames.

**Coherence Measures** Figure 4.12 depicts the coherence measures for the *frame keywords* and BERTopic topics respectively. The number of BERTopic topics was reduced to 14 by initialising the model with `nr_topics=14` to enable comparison. While the averages of the coherence measures are close (0.589 and 0.639) the variance of the reduced topics is a lot higher. This is much in line with subjective evaluation of BERTopic titles in Figure 4.12 and *frame keywords* in Table 4.4. While the most coherent title *2_moon_space_apollo_lunar* seems to unequivocally define a topic, the *frame keywords* for capacity and resources are less clearly defined. In fact, it is likely that the high coherence of *frame keywords* is in part due to the manner they were created. Recall, we selected words corresponding to the highest coefficients of a linear regression model. These coefficients would be high for words frequently occurring in documents with the respective frame, while not occurring as much in documents with other frames. It is probable that many predictive words would co-occur frequently while not being inherently related, resulting in a high coherence measure but relatively low perceived coherence. I.e., creating topics using a linear model trained to predict frames indirectly optimises the coherence measure while not optimising for perceived coherence.

## 4.5 Unsupervised Frame Detection

Extracting narratives in the manner described in Section 3.4 initially generates 991 344 narratives, the distribution of which is plotted in Figure 4.13. The distribution is heavy-tailed, with most narratives only occurring once. The average number of occurrences plotted in red is thus only 1.4. Filtering methods based on the number of occurrences are as such extremely effective. After filtering, I qualitatively analyse the extracted unsupervised frames.

**Filtering** As described in Section 3.4, the unsupervised frame detection performed in this thesis relies on the filtering operations illustrated in Figure 4.14. After extracting narratives, all narratives that occur less frequently than 10 times globally are discarded. This already removes around 95% of all tuples, although it leaves 5 598 which are still too many to examine manually. Based on assumption i) I further require that the tuples occur in at least 50 documents to constitute a relevant frame, which removes a further 90%. Discarding frames that occur too frequently,

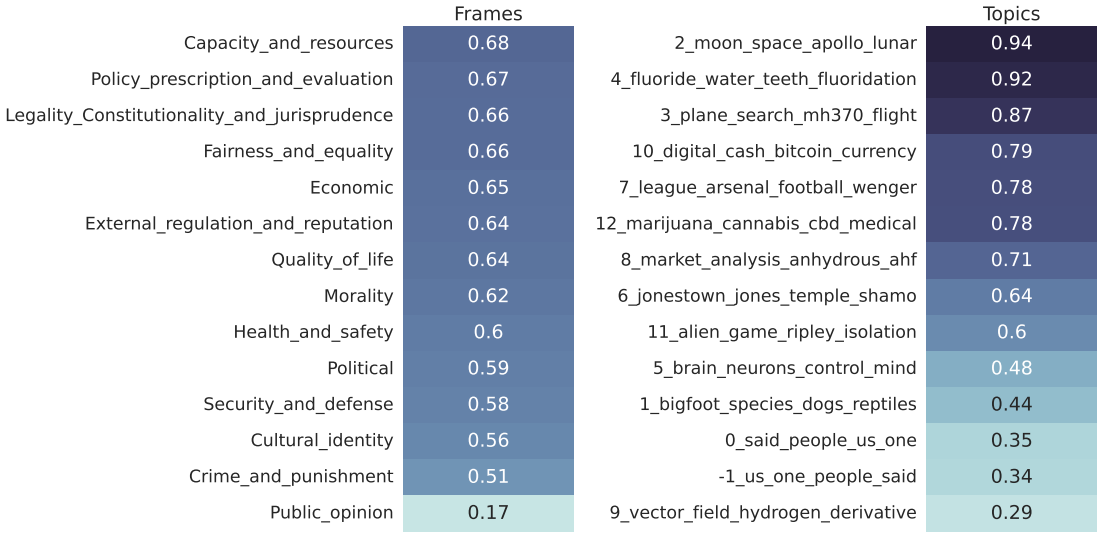



Figure 4.12: Coherence measures for the 30 most significant predictors (words) of frames extracted with linear models (left) and for the 14 reduced topics extracted with BERTopic. Mean coherence is 0.589 and 0.639 respectively. While the means are similar, the variance for frame-topics is much higher.

i.e., in more than 10% of the documents only removes 9 frames such as *government-organization govern* or *person have-org-role*. Requiring relevant frames to be present at least twice in at least 10 documents finally reduces the number of frames to a manageable amount. Filtering out frames that have either positive or negative sentiment results in the 16 frames listed in Table 4.5. All thresholds were determined empirically to strike a balance between removing too many and thus relevant frames and removing too few frames.

**Interpretation of Unsupervised Frames** The frames listed in Table 4.5 are mostly of negative sentiment and strongly relate to the topics occurring in the corpus. Frames marked with a † occur more frequently within the conspiracy sub-corpus at a significance level of 0.01. The hypothesis test was performed using `scipy.stats.fisher_exact` (Virtanen et al., 2020) i.e., $\mathcal{H}_0$ is that the odds ratio of the mainstream and conspiracy populations is one. The conspiracy sub-corpus uses more unsupervised frames overall and emphasises narratives relating to *terrorism*, *fighting* or *rebelling*, and *payment* to the government and not further specified

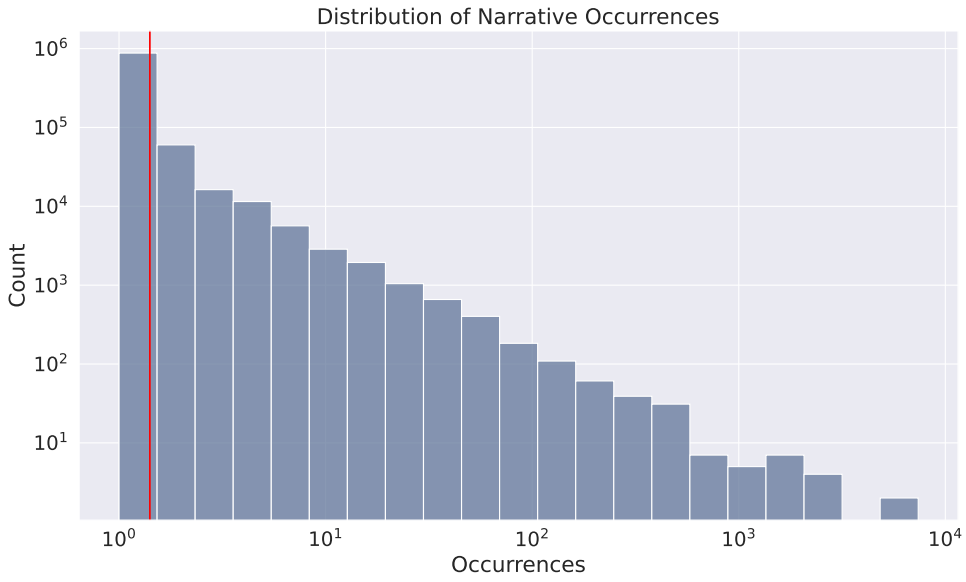



Figure 4.13: Distribution of the number of narratives occurring $x$ times with the average (1.4) marked in red. The distribution is a power law distribution, with most frames only occurring once.

entities. There are no frames that occur significantly more often in the mainstream sub-corpus. The most occurring narrative is *disease cause die*, and it relates to the topics COVID-19 and HIV/AIDS. However, as illustrated by Figure 4.15 the narrative occurs in different contexts. In conspiracy media, it frequently occurs in relation to the terms *vaccine*, *children*, and *cancer* whereas in mainstream media the context is centred on the diseases themselves and terms include *disease*, *AIDS*, *HIV*, *COVID*, and *case*. In the below quote a more specific narrative *measles cause die* occurs in the segment "no deaths in the U.S. from measles" however, in this context the narrative reverses its meaning due to the negation. Instead of being used factually, it is used to emphasise the relative danger of vaccines.

> Since 2005 (and even before that), there have been no deaths in the U.S. from measles, but there have been 86 deaths from MMR vaccine – 68 of them in children under 3 years old. (Miani et al. (2021))

This demonstrates that even seemingly objective narratives can be used to frame issues from very different perspectives. However, it also shows that a simple subject-

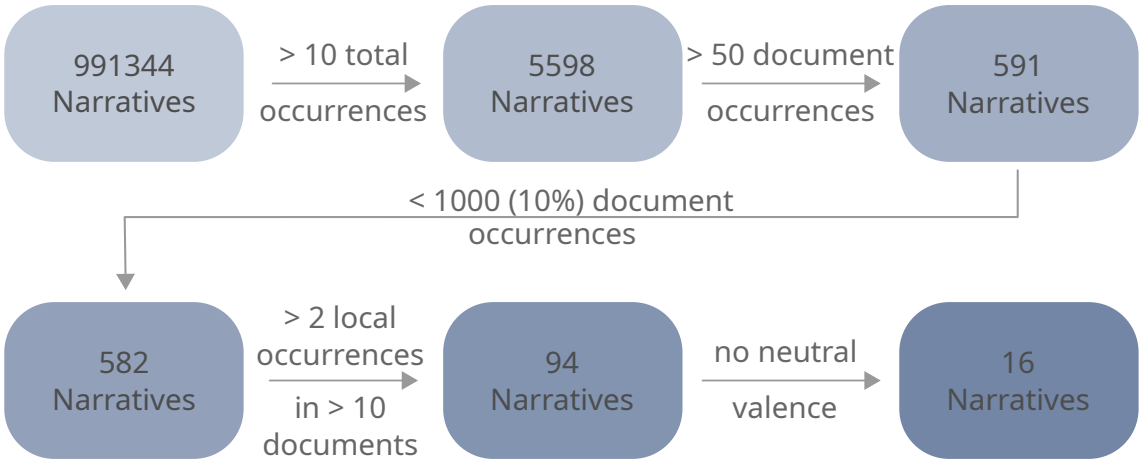



Figure 4.14: The narrative filtering pipeline and the number of narratives remaining after each operation.

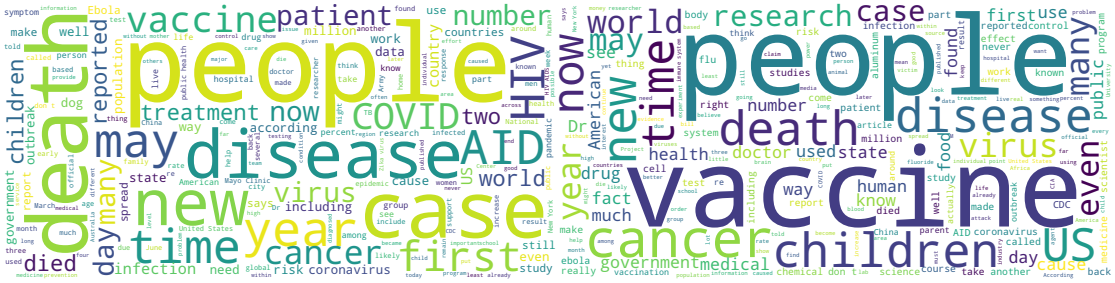



Figure 4.15: Word clouds for the narrative *disease cause death* in mainstream (left) and conspiracy (right) media. In mainstream media the main focus is the disease itself, whereas conspiracy media focuses more on the dangers of vaccines.

predicate-object triple is not sufficient to capture a frame entirely.

Other frames are not as easy to interpret. The *i thank you* frame occurs mostly in direct quotations or dialogues, such as

> "Thank you!" Kunickis wrote from an email address... (Miani et al. (2021))

> James Corbett: Thank you very much for having me on. It's a pleasure. (Miani et al. (2021))

and closing lines of blog posts

| Frame | Sentiment | % | Frame | Sentiment | % |
|---|---|---|---|---|---|
| disease cause die | -0.60 | 3.72 | terrorist attack † | -0.83 | 1.75 |
| person criticize | -0.38 | 2.64 | person protest | -0.25 | 1.66 |
| person play | 0.34 | 2.52 | person rebel † | -0.15 | 1.33 |
| person pay tax † | -0.10 | 2.11 | you risk | -0.27 | 1.05 |
| person fight † | -0.38 | 2.02 | person shoot | -0.34 | 1.02 |
| person prosecute † | -0.40 | 1.86 | i thank you † | 0.36 | 1.02 |
| vaccine effective | 0.48 | 1.82 | person fight fire | -0.61 | 0.70 |
| person pay † | -0.10 | 1.81 | tooth decay | -0.40 | 0.52 |

Table 4.5: Frames extracted using narrative analysis and filtering operations. † narratives occur statistically significantly more frequently in conspiracy media ($P$-value $< 0.01$). There are no narratives that occur significantly more often in mainstream media.

> And as always, stay up-to-date by subscribing to the #NewWorldNextWeek RSS feed or iTunes feed. Thank you. (Miani et al. (2021))

It should therefore perhaps not be considered as a frame as such. The higher relative number of this narrative in conspiracy media however suggests that these media make more use of dialogue and personal writing styles.

Finally, frames like *person shoot*, and *terrorist attack* are bounded and mostly have a concrete topic, in this case public shootings, and 9/11. However, these narratives often have distinct word co-occurrences within mainstream and conspiracy media. Conspiracy media focus more on intelligence agencies and terms like *CIA* or *FBI* occur more frequently.

**Top-Level Narratives** It is vital to recall that narratives extracted using the above-defined method do not always tell the whole story. Consider the following narratives extracted from the last paragraph of a mainstream article: *you care thing, you eat thing, you need concern, you need concern you, thing concern you, nothing safe, test cause safe, you trust test.* Especially the last three narratives *nothing safe*, *test cause safe*, and *you trust test* paint a very alarming image and seem to strongly emphasise the importance that you trust the test.

Now compare the top-level narrative i.e., the narrative we construct by combining and interpreting the atomic narratives with the actual paragraph (own italics):

> Be careful what you eat, its not just Bats you need to be concerned about it seems. With goats, sheep and fruit also testing positive, nothing is safe. *Assuming* you trust the test. (Miani et al. (2021))

Note that assuming refers back to the beginning of the article:

> President Magufuli says tests were found to be faulty after goat, sheep and pawpaw[fruit] samples test positive for COVID-19.
>
> Quite an entertaining example of why these COVID19 test kits should be largely dismissed. (Miani et al. (2021))

with the corresponding narratives *person [Magufuli] say find fault test* and *thing cause recommend dismiss kit*. This new information causes us to entirely revise our top-level narrative. We can dismiss our initial doom and gloom narrative in favour of one that is satirical. This suggests that, while narratives do yield a lot of new information, it would be valuable to examine combinations of frequently co-occurring narratives i.e., top-level narratives. This however escapes the scope of this thesis.

### 4.5.1 Contrasting Unsupervised Frames and Frames

Figure 4.16 depicts the mutual information calculated according to Equation 3.9 between unsupervised frames and the frames defined in the PFC. The matrix is doubly sorted according to the unsupervised frames' (top to bottom) and the frames' (left to right) total mutual information. Unsupervised frames with high MI convey more information about frames than those with low MI and could therefore be interpreted to be more frame-like. Indeed, it is easier to assign PFC frames to the narratives with higher MI. The *person pay tax* narrative has the highest MI for the frames *economic* and *political*, which aligns with the frames we would have intuitively assigned. Narratives towards the bottom of Figure 4.16 such as *you risk* or *person play* are much harder to place. It is however difficult to say whether any of the unsupervised frames constitute new frames in and of themselves. In some sense, everything is framing and one thing can mean two entirely opposite things in different contexts. Generally, narratives are more concrete than the frames from the PFC and can provide additional information on the context of frames. Therefore, they should be considered as a separate but complimentary category.

Equivalently, we can examine the columns of the MI matrix. The higher MI of frames on the left indicates that they are better defined by their unsupervised counterparts. The order of the columns is in part determined by the probability of

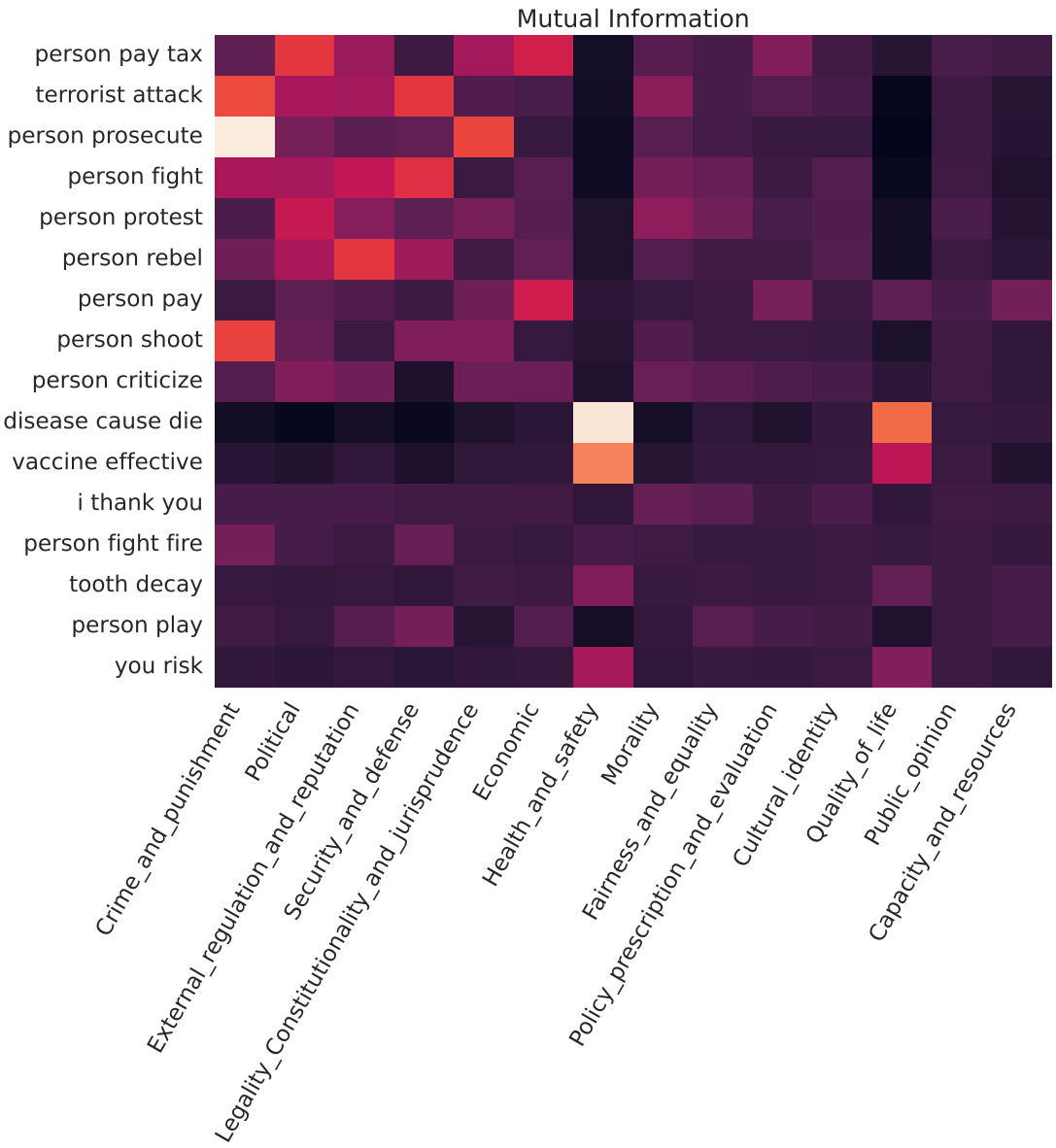



Figure 4.16: Mutual information of frames and narratives. The matrix is doubly sorted according to the rows' and columns' total information. Narratives with high mutual information (sum of rows) are more frame-like.

observing a specific frame and hence the frame count due to the joint probability term $\mathcal{P}(x, y)$ in Equation 3.9. It is also possible, that narratives discarded by the filtering process may have raised the MI of frames towards the right side of the matrix. While this limits definitive conclusions, we can say that the frames on the right are not well defined by the narratives, given the filtering operations. In contrast, frames on the left are strongly supported by the extracted narratives.

In conclusion, we can say that frames extracted in an unsupervised manner via narrative analysis are not equivalent to PFC frames. Whereas some narratives are very related to frames, others are more difficult to place. Due to their atomic nature, they also portray a degree of similarity to topics in that they lack context. As such, narratives are not new frames, but do support existing frames by embedding them into a story.

# Chapter 5

# Conclusion

Building on work done in Ertl et al. (2023), I expand on the theory of contrastive learning and show the effectiveness of contrast sampling as an addition. Contrastive learning can greatly increase performance on tasks with small datasets and is able to generalise well to zero-shot settings within NLP. Furthermore, the contrastive learning paradigm is applicable to multi-label NLP problems. In particular, it is the combination of multiple deep learning techniques that affords superior performance. Contrastive learning, multilingual pre-training and transfer learning, and contrastive sampling are equally important to learning representative language embeddings. The approach was demonstrated on the SemEval (Piskorski et al., 2023) moral framing challenge, winning the zero-shot language Spanish and coming in top-10 on all other languages.

In addition to presenting the multilingual contrastive learning pipeline *mCPT*, I build upon analysis conducted in Ertl et al. (2023) regarding the learned embedding space as well as the generated embeddings. The transformation of the embedding space by applying contrastive training depends on the specific loss function. Contrastive losses that use the cosine similarity to determine the similarity of embeddings organise the embeddings of samples with the same label vector along axes drawn from the origin. In contrast, contrastive losses that calculate the Euclidean distance between embeddings cluster embeddings corresponding to label vectors with high magnitudes around a common centre. However, Euclidean distance-based losses do not perform as well as those that employ cosine similarity. As shown in Wang and Isola (2020), multi-label contrastive learning with losses based on cosine similarity also optimises for uniformity and alignment. However, in multi-label

settings, the dimension of the embedding space must be at least as large as the dimension of the label vector since every unique label vector is assigned its own axis and proximity of axes should reflect similarity of label vectors.

Regarding the difference between frames and topics: topics are more tangible, i.e., more specific. There is no direct mapping from topics to frames, and many frames detected for certain topics are not immediately obvious. Frames always depend on a concrete context, as that is what they are in essence. Frames are the embedding of factual data into a context, for it is context that makes facts say one thing or another. For example, sarcasm can completely reverse the meaning of a factual statement. It is likely for this reason, that more frames are picked up on in highly emotional topics. An appeal to emotions provides some form of context, such as the standpoint of the author. Finally, the lower coherence scores of *frame keywords* compared to topics is a further indicator for lack of specificity.

As with topics, narratives are quite different to frames. Simple subject-predicate-object triples lack context, as we saw in the *top-level narratives* example. As such, base-narratives do not constitute frames per se. Therefore, it would be wrong to say that these narratives represent new frames. However, they are useful for the interpretation of frames in that they add low-level narrative information which can facilitate interpretation. As for the use of narratives, conspiracy media seem to make increased use of narratives that obey the assumptions used to define the filtering criteria. That is, conspiracy media make more use of specific narratives to hammer home a message.

On the whole, both topics and narratives speak for the diversity of frames and show that frames are not reducible to either. Simultaneously, both topics and narratives add further dimensions to textual analysis. In fact, their combination with frames may help realise otherwise overlooked insights, as it allows a more fine-grained analysis.

# Chapter 6

# Reflections and Future Work

This thesis is not exhaustive and offers multiple avenues for future work, both on fundamentals and on the more theoretical side. On the one hand, language models as well as their training procedures can greatly be improved. Specifically, a more theoretically grounded approach for multi-label contrastive learning in NLP might provide insight into further optimisations. Furthermore, while unsupervised learning has seen a lot of use in computer vision tasks and even in some NLP challenges, its application for moral framing analysis remains unsolved. The subjective nature of moral frames makes generation of negative pairs, as required by contrastive learning, challenging. On the other hand, related to the previous point, the open question as to the precise aspects of frames makes defining requirements and necessary assumptions for framing models difficult.

This work also has limitations with respect to the training pipeline itself. As briefly discussed in Ertl et al. (2023), while *mCPT* performed well within the SemEval challenge, winning one of the zero-shot languages, its average placement was 6.2. In contrast, the best team scored an average placement of 2.6. Additionally, the relatively small transformer model employed in *mCPT* doesn't allow for a direct comparison between some more computationally intensive state-of-the-art models. It is possible that a larger language model and batch size would have benefited frame analysis. Finally, annotating additional datasets for evaluations would increase confidence in the model as a whole.

While a part of this thesis focuses on the nature of frames extracted with *mCPT*, what it is precisely that the model picks up on remains undefined. Furthermore, their subjective nature makes frames ephemeral in some sense. The impossibility

of defining frames in such a way that nothing is up to interpretation, makes performance dependent on the applied measure when no labelled dataset is given. Topic and narrative analysis only add context to frames, while not fully specifying them. As such, expert validation might help pinpoint weaknesses of the model.

While narratives serve as a useful comparison to frames, they are not itself without limitations. The assumptions the filtering operations for unsupervised narrative analysis are based on have no theoretical foundations. As such, some relevant narratives are likely discarded. The requirement that narratives occur multiple times per document is especially restrictive, as a semantically equivalent narrative might be formulated in multiple different ways. The construction of top-level narratives and combinations with additional textual features such as conventional frames could also help capture the entire context and thus constitute a more complete frame. Furthermore, analysis on additional datasets could present a larger picture.

# Bibliography

Ali, M. and Hassan, N. (2022). A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186.

Boydstun, A. E., Card, D., Gross, J. H., Resnik, P., and Smith, N. A. (2014). Tracking the development of media frames within and across policy issues. In *ASPA Annual Meeting*.

Boydstun, A. E., Gross, J. H., Resnik, P., and Smith, N. A. (2013). Identifying media frames and frame dynamics within and across policy issues. In *New Directions in Analyzing Text as Data Workshop, London*.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. (2013). API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Burscher, B., Vliegenthart, R., and de Vreese, C. H. (2016). Frames beyond words: Applying cluster and sentiment analysis to news coverage of the nuclear power issue. *Social Science Computer Review*, 34(5):530–545.

Card, D., Boydstun, A., Gross, J. H., Resnik, P., and Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Chuang, C.-Y., Robinson, J., Lin, Y.-C., Torralba, A., and Jegelka, S. (2020). Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of naacL-HLT*, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers):4171–4186.

Entman, R. M. (1993). Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Entman, R. M. (2007). Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.

Ertl, A., Reiter-Haas, M., Innerebner, K., and Lex, E. (2023). Minicpt at semeval-2023 task 3: Multi-label-aware contrastive pretraining for framing prediction with limited multilingual data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Fong, A., Roozenbeek, J., Goldwert, D., Rathje, S., and van der Linden, S. (2021). The language of conspiracy: A psychological analysis of speech used by conspiracy theorists and their followers on Twitter. *Group Processes & Intergroup Relations*, 24(4):606–623.

Gao, T., Yao, X., and Chen, D. (2022). SimCSE: Simple Contrastive Learning of Sentence Embeddings.

Goodman, M. W. (2020). Penman: An open-source library and tool for amr graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 312–319.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2022). Unsupervised Dense Information Retrieval with Contrastive Learning.

Jones, M. D. and McBeth, M. K. (2010). A Narrative Policy Framework: Clear Enough to Be Wrong?: Jones/McBeth: A Narrative Policy Framework. *Policy Studies Journal*, 38(2):329–353.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Liu, S., Guo, L., Mays, K., Betke, M., and Wijaya, D. T. (2019). Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 504–514.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

Miani, A., Hills, T., and Bangerter, A. (2021). LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, 54(4):1794–1817.

Mulder, M., Inel, O., Oosterman, J., and Tintarev, N. (2021). Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 478–488, Virtual Event Canada. ACM.

Piskorski, J., Stefanovitch, N., Da San Martino, G., and Nakov, P. (2023). Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Reiter-Haas, M., Klösch, B., Hadler, M., and Lex, E. (2023). AMR-based Framing Analysis of Health-Related Narratives: Conspiracy versus Mainstream Media. *Under review.*

Reiter-Haas, M., Kopeinik, S., and Lex, E. (2021). Studying moral-based differences in the framing of political tweets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 1085–1089.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai China. ACM.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, 49(1):103–122.

Su, X., Wang, R., and Dai, X. (2022). Contrastive Learning-Enhanced Nearest Neighbor Mechanism for Multi-Label Text Classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.

Tourni, I., Guo, L., Daryanto, T. H., Zhafransyah, F., Halim, E. E., Jalal, M., Chen, B., Lai, S., Hu, H., Betke, M., Ishwar, P., and Wijaya, D. T. (2021). Detecting Frames in News Headlines and Lead Images in U.S. Gun Violence Coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4037–4050, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., and Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Tversky, A. and Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Behavioral decision making*, pages 25–41. Springer, Boston, MA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wang, R., Dai, X., et al. (2022). Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679.

Wang, T. and Isola, P. (2020). Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *International Conference on Machine Learning*, pages 9929–9939.

Weinberger, K. Q. and Saul, L. K. (2009). Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of machine learning research*.

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. (2021). What Should Not Be Contrastive in Contrastive Learning.

Zheng, L., Xiong, J., Zhu, Y., and He, J. (2022). Contrastive learning with complex heterogeneity. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2594–2604.