Moral Embeddings: A closer look at their Performance, Generalizability and Transferability

Dragos-Paul Vecerdea[1] , **Pradeep Murukannaiah** , **Enrico Liscio**
[1]TU Delft

Abstract

Moral values are abstract ideas that ground our judgements towards what is right or wrong. However, with the rapid unfold of moral rhetoric on social media, it becomes increasingly important to place these ideas in a moral frame, contain their harmful effects, and recognise their positive ones. So far, estimating values from opinionated text has posed a challenge due to values' abstract and subjective nature. However, with the latest developments in Natural Language Processing (NLP), we foresee an opportunity to align the study of morality in text with state-of-the-art NLP architectures. Recently published, the Moral Foundations Tweeter Corpus is a milestone in moral classification tasks by offering a dataset that allows for a closer look into how people express moral narratives in social media. In the downstream process of a text classifier, embeddings convert words and sentences into meaningful vectors. Pre-trained on large corpora, they can be fine-tuned, and domain adapted. This study proposes a refinement model, starting from the available dataset, that learns to capture moral information in Sentence-BERT embeddings by applying a state-of-the-art supervised method (triplet loss). We further demonstrate how the refined embeddings improve the accuracy of moral classifiers. Finally, with an improvement of 5% F1-score over models that use pre-trained embeddings, we pave the way towards a generalisable and transferable set of moral embeddings.

Keywords: Morality, Natural Language Processing, Embeddings, Generalizability, Transferability, Sentence-BERT

1 Introduction

Personal values are the abstract motivations that drive our opinions and actions. Shaped by culture and personal experience, they now travel freely in the digital space. For example, studies reveal that moral rhetoric expressed on social media embeds different values depending on the political orientation [1] and can even anticipate violent protests [2]. A better understanding of how we express these moral perspectives and how they vary across groups (e.g. political, cultural, religious) would enable us to take action on harmful ideas and eventually design Artificial Intelligence (AI) aligned with our virtues. The recent improvements of Natural Language Processing (NLP) techniques allow us to design models that accurately estimate individual values from the opinionated text. However, not performed with the (current) state-of-the-art models (e.g. BERT [3], MPNet [4]) and heavily reliant on feature engineering [5], existing work reveals the challenges posed by the subjective nature of the task: agreement on a universal set of values, data annotator's subjectivity.

Moral Foundations Theory [6] is a widely adopted theory in social psychological studies. It proposes five innate and universally available moral foundations: *care*, *authority*, *fairness*, *loyalty* and *purity*. To hold explainable and comparable findings, we frame our research in the terms proposed by MFT. In support of this choice, we use the Moral Foundations Twitter Corpus [7], a sizeable and adequately annotated dataset. Its richness in topics allows for an extensive analysis of state-of-the-art architectures in the multi-label moral classification task. Widely adopted ([8], [5]), it offers the premises of an in-depth understanding of morality expressed in social media.

Embeddings are mappings of words (or sentences) to numeric representation that feed machine learning models. They are trained on large corpora and learn to project text onto a low dimensional continuous space. The resulting vectorial representations are meaningful as words (or sentences) with similar meaning are close (e.g. Euclidean distance, cosine similarity) to one another in the embedding space. Different techniques are used to re-train these embeddings to capture domain or task-specific information. Thus, their quality has a tangible impact on the model's performance. However, their computation is often not explicit (e.g. BERT), or they are not domain adapted, making the learning process more complex for the model on top. Furthermore, past moral classifiers have either not used stat-of-the-art embeddings [7] or did not train them to learn moral foundations [8], moving the pressure on the model's learning capabilities. Significantly, fine-tuned embeddings' usage is not limited to the model and task they are designed for in the first place. For example, they can reveal hidden social biases [9], and more generally, reveal domain-specific particularities. Our work

successfully embeds morality into these vector representations and studies their direct impact on the moral classifier and ability to transfer learning and generalize across different discourse domains.

To achieve their moralization, state-of-the-art embeddings (SBERT [10]) are fine-tuned using a supervised method called triplet loss [11]. For a complete classification model architecture, on top of these vector representations, two dense layers are added. While fine-tuning proves an expensive operation (≈4min per epoch, entire dataset[1]), training the moral classifier will take few seconds (≈2s per epoch, entire dataset).

This work proposes three multi-stage embeddings fine-tuning methodologies to overcome the identified limitations of related research and obtain a complete technical overview of the proposed work. It successfully leverages the availability of an adequately annotated dataset (MFT labels), pre-trained language models' transferability [12] and supervised tuning methods (triplet loss [11]). Nonetheless, we implement a set of models (LSTM and Bi-LSTM) for an objective comparison with past papers that serve as baseline models for our analysis.

We identify the main contributions of this paper as follows:

- A lightweight method for effectively optimizing the embedding itself, rather than increasing the complexity of the classification model.
- Demonstrate the potential and limitations of transfer learning and generalizability in moral embeddings.
- Proposes a highly accurate alternative to the current state-of-the-art models in moral classification.

The paper is structured in the following way: In section 2 we make an analysis of related work. Section 3 presents the dataset and the proposed embedding fine-tuning methods. Section 4 takes three perspectives on moral embeddings and proposes relevant experiments. In section 5, we reflect on the experiments' results. In section 6, we perform responsible research. Section 7 concludes the work and proposes future improvements.

2 Related work

Moral values are personal and abstract drives that push us to action. The implied degree of subjectivity makes their study virtually impossible without an agreement on a moral framework. Moral Foundations Theory [6] is the common ground for past attempts ([13],[7],[8],[5]) in moral classification. Hoover et al. (2012) define five "irreducible basic elements" (*care*, *authority*, *fairness*, *purity*, *loyalty*) of morality and describe their theory as one "to be expanded" and that gives researchers "a common language for talking about the moral domain".

Early related work, The Morality Machine [14], proposes a suitably annotated dataset (Moral Foundations Theory [6]) but its size (N = 8,292) and narrow topic ("Grexit") are limiting the potential of this work. More recently, a leap forward was made in moral classification with the release of the Moral Foundations Twitter Corpus, our choice . Hoover et al. [7] published the Moral Foundations Twitter Corpus, a dataset of 35 thousand tweets annotated with the same annotation schema and sub-divided into seven smaller sub-datasets with topics ranging from the 2016 US presidential elections to Hurricane Sandy. The Twitter Corpus paper, however, does not explore, in the proposed classification models, state of the art embeddings (GloVe [15], SBERT [10]) and NLP architectures (e.g. BERT, Bi-LSTM). Further, Araque et al. [5] achieve state-of-the-art results on the proposed dataset and demonstrates how a moral lexicon (*MoralStrength*, 1000 lemmas) can be exploited at the document level using complex feature engineering techniques. However, while demonstrating the potential of morality as a dimension in sentence vector representations, the experiments were limited in complexity and only differentiated between moral foundations and non-morality (binary classification) and not between foundations or even further, between vices and virtues. Moreover, feature engineering limits the models' performances to the quality of the lexicon and it cannot transfer learning to other languages when compared to the state-of-the-art architectures [16]. Luenen [8] attempts to overcome these limitations and successfully proposes two Bi-LSTM models that differentiate between vices and virtues within the moral foundations and rely on pre-trained embeddings (word2vec and BERT). Despite a thorough analysis of the generalizability of such models, fine-tuning sentence or word representations is not explored and achieved results (60% weighted average F1-score, on MFTC dataset, 11 labels) are lower than ours (72%).

Domain adaptation embedding methods as those presented in this papers have made the subject of papers focusing on sentiment analysis and serve as a basis [17]. However, they have not explored the potential of such techniques in moral foundation classification, and none has used the potential of pre-trained SBERT (Semantic Textual Similarity) in conjunction with fine-tuning and lightweight architectures but have rather relied on complex models or feature engineering.

3 Methodology

We aim to recognise moral foundations (Table 1) in opinionated text and distinguish between their *vices* and *virtues*. To achieve this, we define our problem as a multi-label classifier with 11 classes. For each moral foundation, we have two distinct classes (e.g. *care/harm*). Those texts that express no foundation belong to the *non-moral* class. Compared to a multi-class model, ours allows for data entries to have multiple labels at a time.

In the downstream process of moral classification, we focus our attention on word and sentence embeddings. Morality is encapsulated in these representations using fine-tuning techniques, and the model's efficiency increases. Aside, we implement two simpler architectures: LSTM and Bi-LSTM to reproduce and compare our work to past moral classifiers.

Formally, we can define our goal as follows: given a set of texts $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$, an embedding space $\mathcal{E}$ and a classification model $\mathcal{M} : \mathcal{E} \mapsto \mathcal{C}$, we wish to learn a mapping $f : \mathcal{T} \mapsto \mathcal{E}$, for which $\mathcal{M}$ performs best (refer to section 3.5 for a definition of performance metrics).

[1] Using the computation recources presented in Section 6.2

Throughout the paper, we analyse the effectiveness of the proposed moral embeddings from multiple technical perspectives (refer to section 4):

- **Performance**: how much do fine-tuning embeddings (triplet loss) increase the performance metrics of the moral classifier (compared to pre-trained embeddings)
- **Generalizability**: are fine-tuned embeddings suitable for moral classification on a different dataset (than the one used to train the embeddings)
- **Transferability**: can fine-tuned embeddings replace pre-trained embeddings in moral classification

Next, we present the critical components of the methodology. The order follows a bottom-up approach, similar to a complete classification pipeline, starting with data, followed by its numerical representation (embeddings) and eventually discussing the models and their evaluation.

3.1 Moral Foundations Theory

First proposed by Haidt et al. (2004), The Moral Foundation Theory serves as a common ground for analysing people's morality. In its earliest development, it focused on the study of moral values within political ideologies. Later, it became a universal moral frame that offered people a common "language" to analyse morality. The theory defines a set of 5 foundations (Table 1) that stay at the base of our choices and actions and that we all hold to a degree. These foundations have manifestations that individuals can perceive as positive (virtue) or negative (vice). However, most attempts in moral foundations classification are not differentiating between the vice and virtue of a foundation. Increasingly often, ethical dilemmas arise, and polarisation takes various shapes in social media. In this context, it is especially relevant to understand better what people perceive as praiseworthy (virtue) or blameworthy (vice). Despite the additional technical challenges (11 classes), we consider this insight relevant to understanding moral narratives.

3.2 Dataset

Structured into seven subsets (Table 2), the Twitter Corpus benefits from many annotators (up to 8 per tweet), a considerable size (more than 35 thousand tweets) and the diversity of ideas in social movements (e.g. politics, human rights, natural disaster). Together they form a complete view of how morality takes shape in social media, and this has transformed it into a benchmark for multi-label moral value classification.

Data aquisition

Data acquisition was initially challenged by the low availability of the original tweets, fetched using the Twitter API. Only 49,9% of the total data was publicly available at the time, and two datasets were entirely unavailable (MeToo and Davidson). Consequently, we requested the dataset from the authors of the Moral Foundation Twitter Corpus paper. Kindly, they offered us the whole dataset. Experiments to follow were all conducted using the dataset in its entirety. With that, we hope to achieve an objective comparison to past work.

Foundation	Definition
Care Harm	This foundation is related to our long evolution as mammals with attachment systems and an ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.
Fairness Cheating	This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy
Loyalty Betrayal	This foundation is related to our long history as tribal creatures able to form shifting coalitions. It underlies virtues of patriotism and self-sacrifice for the group. It is active anytime people feel that it's "one for all, and all for one."
Authority Subversion	This foundation was shaped by our long primate history of hierarchical social interactions. It underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
Purity Degradation	This foundation was shaped by the psychology of disgust and contamination. It underlies religious notions of striving to live in an elevated, less carnal, more noble way. It underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants (an idea not unique to religious traditions).

Table 1: Definitions of moral foundations proposed in MFTC [6]

Corpus	Description
All Lives Matter	Tweets related to the All Lives Matter movement
Black Lives Matter	Tweets related to the Black Lives Matter Movement
Baltimore Protests	Tweets posted during the Baltimore protests against the death of Freddie Gray
2016 U.S. Presidential Election	Tweets posted during the 2016 U.S. Presidential Election
Hurricane Sandy	Tweets related to Hurricane Sandy, a hurricane that caused record damage in the United States
#MeToo	Tweets related to the #MeToo movement
Davidson Hate Speech	Tweets collected by Davidson et al. (2017) for hate speech and offensive language research

Table 2: Moral Foundations Twitter Corpus (MFTC) Discourse Domains [7]

Preprocessing

We closely analysed the raw data and designed six preprocessing pipelines adequately adapted to different embeddings. Notably, most Tweets include '#' symbols, usernames, URLs, concatenated words, abbreviations and spelling mistakes. To fix spelling mistakes, remove unnecessary tokens (e.g. URLs, numbers, dates) and split merged words, we use the ekphrasis[2] library [18], a collection of lightweight text tools geared towards text from Twitter.

Different models and embeddings are expected to behave more efficiently with different forms of preprocessing. Therefore, we introduce a set of methods common to all models:

- Lowercase
- Remove usernames, URLs and numbers
- Remove # symbols
- Remove punctuation
- Remove emojis
- Remove stopwords

[2]https://github.com/cbaziotis/ekphrasis

Further, we differentiate between context-dependent and context-independent embeddings. While articulated words can change the semantic meaning of context-dependent sentence representation, for context-independent word representations, empirically (Figure 1 and 2), the model performs better when words are lemmatised and stop words are removed (the vocabulary is reduced).

We extend the list of preprocessing methods with more complex operations:

- Lemmatization
- Word segmentation
- Spell correction
- Translate emojis to words

We crafted six strategies (Table 3), using the introduced methods. However, it proved infeasible to perform complex experiments with six different processing strategies. To overcome this, we run the baseline models with pre-trained embeddings for all strategies (Figure 1 and 2) on the entire MFTC dataset. We then analysed for which one does the model have the highest F1-score. Finally, two choices were made, each supporting a different architecture. Strategy 3 is the choice for LSTM and Bi-LSTM and Strategy 5 for SBERT. For SBERT, four strategies are performing very similarly, and we choose the most complex and technically adequate for context-dependent embeddings.

Strategy	S0	S1	S2	**S3**	S4	**S5**
Lowercase	✗	✓	✓	✓	✓	✓
No usernames, urls and numbers	✗	✓	✓	✓	✓	✓
No # symbol	✗	✓	✓	✓	✓	✓
No punctuation	✗	✓	✓	✓	✓	✓
No emojis	✗	✓	✓	✗	✓	✗
No stopwords	✗	✗	✓	✓	✗	✗
Lemmatization	✗	✓	✓	✓	✗	✗
Word segmentation	✗	✗	✗	✓	✓	✓
Spell correction	✗	✗	✗	✓	✓	✓
Emoji to words	✗	✗	✗	✗	✗	✓

Table 3: The six pre-processing strategies, varying in complexity. Strategies in bold are the final choices for the experiments. 3rd is the choice for word-embedding models and 5th is the choice for sentence-embedding models

Annotations

At least three annotators annotate each tweet with one of the ten moral sentiments (or non-moral). For consistency reasons, we apply a majority vote on the annotations. This process is followed by labelling as non-moral those sentences for which no majority vote exists. After applying these operations, some tweets are labelled both as *non-moral* and *moral*. Our option is to prefer false positives to false negatives concerning moral labels. Therefore, the heuristic choice is to remove the non-moral label for those tweets (around 3% of the total) and consider them moral. Table 4 shows the final distribution of labels.

An immediate concern is a sheer imbalance towards the *non-moral* class, present across all datasets. The models' learning process may be harmed by this bias and lead to overfitting. To mitigate these risks, we propose a balancing strategy: the *non-moral* class is randomly downsampled in train data until it reaches a cardinality equal to the second-most

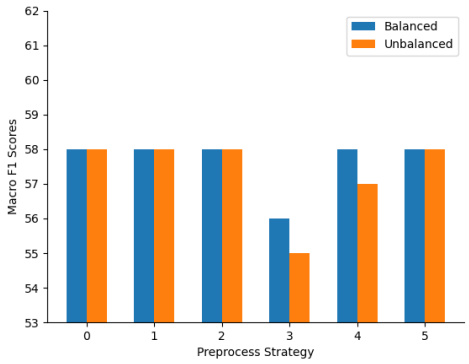

Figure 1: RNN with SBERT. Moral values classification on entire dataset. 1st, 2nd and 5th strategy perform best

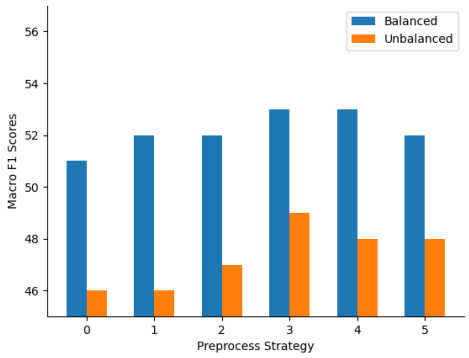

Figure 2: LSTM with GloVe. Moral values classification on entire dataset. 3rd strategy performs best

	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
subversion	91	257	303	7	165	874	451
authority	244	17	276	20	169	415	443
cheating	505	519	876	62	620	685	459
fairness	515	133	522	4	560	391	179
harm	735	244	1037	138	588	433	793
care	456	171	321	9	398	206	992
betrayal	40	621	169	41	128	366	146
loyalty	244	373	523	41	207	322	415
degradation	122	28	186	67	138	941	91
purity	81	40	108	5	409	173	56
non-moral	1744	3848	1583	4509	2502	1618	1313

Table 4: Number of labels per dataset. Majority vote (>= 50%) annotation strategy. Note: this is calculated for the raw data. For some pre-processing strategies, few tweets are dropped.

often class. In contrast, the test data is not downsampled. To quantify the unbalancedness of each dataset and the expected impact of the strategy mentioned above, we calculate the **Shannon Entropy** (normalised):

$$Entropy = -\frac{\sum_{i=1}^{k} \frac{c_\text{i}}{\sum_{i=1}^{k} c_\text{i}} * log(\frac{c_\text{i}}{\sum_{i=1}^{k} c_\text{i}})}{log(k)}$$

$k =$ number of classes

$c_\text{i} =$ cardinality of i$^{\text{th}}$ class

	Entropy	
Dataset	Balanced	Unbalanced
ALM	0.88	0.80
Baltimore	0.85	0.59
BLM	0.90	0.88
Davidson	0.80	0.17
Election	0.93	0.80
MeToo	0.94	0.91
Sandy	0.89	0.87
All	0.96	0.80

Table 5: Entropy of label distributions for balanced and unbalanced datasets. Closer to one means balanced

Table 5 shows how class entropy considerably reduces when balancing the dataset. However, *Davidson* and *Baltimore* are particularly imbalanced and the *non-moral* label accounts for more than 90% of the total. For these datasets, we consider that the proposed balancing strategy might not suffice. After an empirical analysis, we concluded that the cause for the extreme imbalance is the lack of majority vote agreement and a small number of annotators (three or four). We will consider these aspects when reflecting on the results.

A showcase of how the proposed balancing strategy can affect the model's performances is showed in Figure 1 and 2. As it consistently achieves better F1 scores across the preprocessing strategies, we carry the experiments to follow using the balancing method.

3.3 Embeddings

GloVe

GloVe [15] is an unsupervised learning algorithm for obtaining vector representations for words by leveraging a *word-word-co-occurrence* matrix. The resulting representations achieve better results compared to *word2vec* embeddings (used in [8]) on a number of benchmark tasks (e.g. word analogy). With similar results on a sentiment analysis task [19] and yet not explored in moral classification, our choice for GloVe is also motivated by the availability of a flexible fine-tuning library (unsupervised): Mittens. *Mittens* is "a simple extension of the GloVe representation learning model" [20] that updates pre-trained embeddings with domain-specific data.

Fine-tuning with Mittens is an expensive operation for a dataset (MFTC) with a vocabulary size of more than 20.000 tokens. However, many words ($\approx$ 5900) in the dataset are out-of-vocabulary (pre-trained GloVe has no mapping to a numeric representation). Using Mittens, we trained embeddings only for the OOVs and run the baseline models with the updated GloVe embeddings. Model performance *decreased* and, in consequence, for the baseline models, we will use pre-trained embeddings, and OOVs will be mapped to a special token (zero valued vector).

SBERT

A standard critique of state-of-the-art Language Models (e.g. BERT, RoBERTa, MPNet) is that sentence embeddings are not explicitly computed ([10], [12]). Nonetheless, the averaged output layer of these models produces sentence representations, but this technique yields rather bad results [10] and is limited in utility.

Sentence-BERT [10] is a modified Transformer Network [21] that produces semantically meaningful embeddings and addresses the presented drawback. SBERT adds a pooling operation to the output of (BERT / RoBERTa / MPNet) to derive a fixed-sized sentence embedding. In order to fine-tune these Language Models, siamese and triplet networks are used to update the models' weights such that the produced sentence embeddings are semantically meaningful and can have operations applied on them (e.g. cosine similarity). This architecture (Appendix C, Figure 5) allows efficient semantic similarity search as well as clustering.

Not explored in moral classifiers before, we use embeddings designed for STS (Semantic Text Similarity) tasks, where embeddings of sentences are used to lookup for semantically similar text. This choice allows for the fine-tuned embeddings to be suitable not only for the classification task they are trained for but for further use in STS on moral topics. We continue our experiments using the pre-trained SBERT that achieves highest results in benchmarks: *stsb-mpnet-base-v2* [3]

Triplet loss is the loss function that our SBERT embeddings use to learn moral values. It builds triplets out of the training data, consisting of an anchor sentence, a positive sentence (which has a moral value in common), and a negative sentence (which is not labelled with the shared value). The cost function is computed for each triplet, and the model weights are adjusted to minimise it. This loss function learns the embedding model to cluster together (in the vectorial space) sentence representations that share moral values and separates those that do not:

$$Loss = max(0, Dist(a, p) - Dist(a, n) + margin)$$

$a =$ anchor, $p =$ positive, $n =$negative

3.4 Classification Models

Embeddings produce vector representations of the input text. On top, a classification model is used to learn how these representations correspond to certain classes (MFT).

[3] https://huggingface.co/sentence-transformers/stsb-mpnet-base-v2

NN

Neural networks are a class of machine learning algorithms [22]. Architecturally, they are organised in layers, comprised of artificial neurons that mimic the biology of the brain. Together these neurons form a connected graph with weighted edges. While learning, data travels through the graph in a single direction ("feed-forward") and produce an output. The weights are then adjusted, so the output is closer to the ground truth. A simple NN architecture, composed of two Dense Layers and two Dropout Layers (Appendix C), is used on top of the SBERT embeddings. Its simplicity allows for short training times and proves sufficient to learn morality, given the complex and already semantically meaningful embeddings that feed the model.

LSTM Bi-LSTM

Long short-term memory (**LSTM**) [23] represents a type of recurrent neural network (RNN), that is, a kind of recursive NNs. LSTM architectures are suitable for sequences of data and learn potential dependencies in the sequence. Bidirectional Long short-term memory (**Bi-LSTM**) apply the same learning process as LSTMs but in both directions through the input sequence. In our experiments, these architectures fit with word-level embeddings that, for a sentence, produce a sequence of numeric vectors. In contrast, LSTMs/Bi-LSTMs are unsuitable for sentence embeddings (SBERT), which produce a single vector representation with word dependencies already embedded.

Not current state-of-the-art, both models (Appendix C) using LSTM and Bi-LSTM respectively will be used in conjunction with GloVe embeddings. The choice to include them in this paper is rooted in the pursuit of an objective comparison with [7] and [8].

3.5 Metrics

A suitable choice of an evaluation metric was subject to empirical analysis. Most importantly, we reflected on class imbalance, metrics used by other papers, and how the used models vary to this choice. To correctly reflect the value and potential of the work, we decided upon two metrics:

Generic F_1-score:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Micro Averaged F1-Score

This f1-score metric will account for the unbalancedness of the dataset, present even after the balancing strategy has been applied (Figure 5). More interestingly, some experiments (Table 7) show improvements for this metric, leaving the Macro F1-score unchanged. This proves the relevance of the Micro F1-score in better understanding fine-tuned embeddings in the model's learning process.

Macro Averaged F1-Score

This metric will offer a perspective unaffected by the class imbalance (same weight to all classes). However, it is difficult to anticipate the distribution with which moral foundations occur, and to extrapolate on the available dataset could be misleading (e.g. small size, cultural differences). For that, we add the Macro F1-score to our analysis to eliminate part of the bias brought by the class distribution. Additionally, the Moral Foundations Twitter Corpus paper [7] opts for this metric.

4 Experiments

The three proposed experimental setups will cover three key aspects of the moral embeddings:

- **Performance**: to gather insight into the performance of embeddings, we narrow the focus to a dataset and measure the impact fine-tuning has on the classifier.
- **Generalizability**: we measure how does fine-tuning embeddings on a different dataset than the one they are used in the classification task impact their performance.
- **Transferability**: transferability study adds an additional layer to generalizability: fine-tuning happens twice, first cross-domain, followed by domain-specific training.

Throughout the series of experiments, hyperparameters (Appendix B), the models' architectures (Appendix C) and splits in train and test data (random seeds), remain the same. Thus, the first experiment will be done on LSTM, Bi-LSTM with GloVe and NN with SBERT. The subsequent two experiments will only test for models with the latter architecture. All models will use the balanced dataset and trained to perform multi-label text classification on 11 labels.

4.1 Experiment 1: Models Performance

LSTM and Bi-LSTM

- A target dataset is chosen and 10 folds are created.
- Pre-trained GloVe[4] embeddings are used to train and test the models (refer to 3.4).

SBERT

- A target dataset is chosen and 5 folds are created.
- For each fold (80% train data and 20% test data), embeddings are fine-tuned on train data.
- Fine-tuned embeddings are used to train (on the same 80%) the model (refer to 3.4). It is then tested on the remaining 20%.
- An additional model is trained likewise, using pre-trained embeddings [5], to obtain a baseline for SBERT.

4.2 Experiment 2: Embeddings Generalizability

- A target dataset is chosen and using the proposed fine-tuning method, embeddings are trained on the remaining six datasets.
- The resulting embeddings are used to train and test a model on the target dataset. The model is tested and trained using ten-fold validation.

[4] Wikipedia 2014 + Gigaword 5, 6B tokens, 300 dimensions

[5] stsb-mpnet-base-v2, 728 dimensions

4.3 Experiment 3: Embeddings Transferability

- A target dataset is chosen and using the proposed fine-tuning method, pre-trained embeddings are trained on the remaining six datasets.
- Resulting embeddings are fine-tuned a second time, on 80% of the target dataset.
- Twice fine-tuned embeddings are used to train the classifier(on the same 80%). The model is tested on the remaining 20%.

5 Results and Discussion

GloVe embeddings used with LSTM and Bi-LSTM have offered strong baseline results, but their context-independent nature makes encapsulating morality only possible using feature engineering. Nonetheless, the results remain below the current state-of-the-art, and the embeddings' transferability potential is limited to the model used on top.

SBERT proves to be a robust embedding for all addressed experiment setups. Compared to BERT, it offers similar results [24] but allows for usage in tasks like Semantic Textual Similarity, operations between embeddings (cosine similarity), and clustering [10]. The achieved results present an excellent moral classifier and reveal the potential of the proposed techniques to produce embeddings that could become the standard in moral classification. Nonetheless, the limitations (e.g. class imbalance, annotator's inconsistency) of MFTC reflect in the results.

We present the results and draw the conclusion for all the researched aspects of moral embeddings.

5.1 Performance

Baseline models, LSTM and Bi-LSM, set a high bar in moral classification. They achieve, on the 11 classes multi-label classifier, a Macro F1-score of 53% and 55%, respectively (Figure 6). According to Constantinescu [25], training for moral foundations (6 labels, not differentiating between vice and virtue) achieves similar results as training for all moral values (11 labels) and when testing, ignoring the model's misclassifications within a foundation (e.g. mistake care for harm). Therefore, we only perform our research with 11 label models as this can be easily transformed into a six-label model with no performance loss.

Our baseline SBERT model, a set of pre-trained embeddings (stsb-mpnet-base-v2) with a simple neural network on top, achieves astonishing results. With a 58% Macro F1-score, it shows how a moral classifier can be trained in less than a minute ($\approx$1s per epoch) and have consistent results using SBERT's transferability. We continue our performance analysis and fine-tune the pre-trained embeddings on each dataset (and once on the entire MFTC). Triplet Loss tuning significantly improves the model's F1-scores and achieves a Macro F1-score of 63% next to a Micro F1-score of 72% for the entire MFTC. Performance improvements are consistent across all datasets, and Figure 6 reveals how fine-tuning increases scores by up to 11% (Election and ALM).

5.2 Generalizability

The underlying challenge of this experiment is that the model learns from embeddings that have been fine-tuned on other datasets and implicitly on other moral narratives. By comparing to pre-trained embeddings, for few datasets (ALM, BLM), the fine-tuned embeddings generalize well (Table 7), while for others, pre-trained embeddings achieve better results. The three improved datasets have in common the social polarization (two opposing and extremist ideas) and the values the models learn with the highest accuracy (Table 8) in the Performance setup. This experimental setup, put in perspective of the previous experiment, demonstrate that, on average, fine-tuning embeddings on the target dataset is needed as it leverages the full potential of SBERT.

5.3 Transferability

This experiment produces nearly identical results to the first experiment, with slight improvements for Election and MeToo and a slight decrease (maximum 2%) for the rest. As fine-tuning happens twice, this experiment showcases how SBERT embeddings can be affected by repeated training. Overall, these results are encouraging as second fine-tuning requires a sixth of the time needed by the first, and it offers embeddings suitable for all seven topics. Further experiments should be carried out to understand the catastrophic forgetting [26] these embeddings could suffer after consequent fine-tuning. Nonetheless, the results offer a partial image of how moral embeddings could replace pre-trained ones.

5.4 Discussion

Baseline results (58% Macro F1-score) for SBERT architecture show how these embeddings are suitable for moral classification even when no fine-tuning is applied. Further, the improvement (5%) brought by triplet loss tuning validates its technical adequacy for the studied classifiers. It is particularly interesting to note that the used SBERT embeddings were designed and pre-trained with a different task in mind: Semantic Text Similarity[6] (STS) [27]. Given the evidence that SBERT transfers knowledge from STS to the moral classifier, we postulate a hypothesis: *semantically similar text expresses similar moral values*. We propose an experimental setup to test this, in Section 7.1.

The Generalizability experiment falls short and provides scores below what SBERT embeddings can achieve with no fine-tuning. Therefore, for fine-tuning to improve the model's performance, embeddings' train data must share the same domain of discourse with the test data. Embeddings produced by this experiment (second experiment) are further fine-tuned within the study of transferability (third experiment). Obtained results are close to those produced by the first experiment's methodology. Therefore, two consequent fine-tunings do not affect the quality of the embeddings, and for the presented setup, fine-tuned embeddings could replace the pre-trained ones with no decrease in performance. Nonetheless, to confirm such a hypothesis: *fine-tuned embeddings on*

[6] https://paperswithcode.com/sota/semantic-textual-similarity-on-sts-benchmark

Model	F1-score	All	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
LSTM (GloVe Wiki)	Micro	0.61	0.55	0.59	0.76	0.00	0.58	0.51	0.56
	Macro	0.53	0.47	0.22	0.73	0.00	0.46	0.35	0.32
Bi-LSTM (GloVe Wiki)	Micro	0.60	0.53	0.59	0.78	0.13	0.58	0.53	0.57
	Macro	0.55	0.58	0.29	0.77	0.02	0.50	0.50	0.40
SBERT (pre-trained)	Micro	0.63	0.58	0.63	0.76	0.06	0.61	0.57	0.63
	Macro	0.58	0.52	0.34	0.76	0.02	0.55	0.55	0.48
SBERT (experiment 1)	Micro	0.72	0.69	0.70	0.84	0.93	0.72	0.60	0.65
	Macro	0.63	0.59	0.34	0.83	0.08	0.56	0.56	0.45

Table 6: Baseline models and **Performance** of fine-tuned embeddings for **moral values** classification.

Model	F1-score	**ALM**	Baltimore	**BLM**	Davidson	**Election**	MeToo	Sandy
SBERT (experiment 2)	Micro	0.63	0.62	0.78	0.01	0.62	0.57	0.60
	Macro	0.57	0.30	0.76	0.01	0.52	0.54	0.46
SBERT (experiment 3)	Micro	0.66	0.68	0.83	0.91	0.70	0.59	0.64
	Macro	0.57	0.33	0.82	0.01	0.58	0.57	0.45

Table 7: **Generalizability** and **Transferability** of embeddings for **moral values** classification.

Moral Value	MFTC	ALM	BLM
subversion	0.44	0.22	0.81
authority	0.62	**0.79**	**0.91**
cheating	0.69	0.67	0.87
fairness	0.72	**0.81**	**0.91**
harm	0.68	0.63	0.79
care	**0.75**	0.67	0.77
betrayal	0.44	0.00	0.88
loyalty	0.66	**0.78**	**0.94**
degradation	0.55	0.67	0.73
purity	0.53	0.57	0.76
non-moral	0.82	0.81	0.82

Table 8: F1-score per moral value for MFTC (all datasets), ALM and BLM. **Performance** setup.

MFTC can replace pre-trained embeddings in moral classification more properly annotated data is needed.

Out of the three experiments, the one testing for *Performance* validates its hypothesis with the most convincing results: *morality can be learned more efficiently when SBERT [10] embeddings are fine-tuned with triplet loss [11].* Having achieved a 72% Micro F1-score, we can conclude that SBERT has its place in moral classifiers and its state-of-the-art results in STS can be laveraged succesfully.

6 Responsible Research

In emerging AI technologies, recent history has proved that new and unexpected ethical dilemmas arise. This research not only falls within reach of the AI field but is also preoccupied itself with seeking people's morality, embedded in the written text travelling across digital platforms.

In the realm of highly complex AI architectures that could potentially learn to accurately recognize people's values, we identify a list of points of attention:

- Ethical Considerations
- Reproducibility

These topics are discussed with the purpose of containing the negative social impact of the research and support academia in pursuing further research of the proposed classification task.

6.1 Ethical considerations

"Engineering ethics is professional ethics, as opposed to personal morality. It sets the standards for professional practice" [28]. To ensure ethical engineering, we reflect on the impact of the developed software.

The limited availability of suitably annotated datasets restricts our work. The reported results reveal the potential of NLP (Natural Language Processing) models to learn morality in the text but have only exploited a small collection of Tweets. In the future, we expect more datasets to become available as a public or private resource. Therefore, we are in the early stages of understanding whether our techniques extend to more contexts (e.g. politics, public debates, school, other social media platforms). Achieving universality for moral NLP models would be both a breakthrough and a social responsibility. It is early to predict how one would use such a tool, but we can reflect on a twin domain (Computer Vision) and the ethical challenges it faces.

For example, facial recognition technology (Computer Vision) can expose individuals' political [29] or sexual orientation [30]. These technologies raise a concern about a possible decline of privacy. Similarly, moral values can predict political orientation [1] and there is only one step left to bridge the gap: identify values in text. To reduce risks, we recommend the implementation of an ethical framework in future academic work. In a product or application using the moral classifier, such a framework could be insufficient [31], and an adequate ethics risk assessment metric system can be used [32] in addition.

Nonetheless, we should raise awareness of the possible hidden bias of the data (Moral Foundations Twitter Corpus) as it lacks demographics information such as age, sex or ethnic group [33]. This is a barrier in understanding whether the used models have developed worrying biases. In contrast, data protection is no concern yet for this work. No trace of identification was left after pre-processing (removed username).

6.2 Reproducibility

We define **reproducibility** in empirical AI research as "the ability of an independent research team to produce the same results using the same AI method based on the documentation made by the original research team" [34]. Achieving it has become a pillar for scientific progress and increases the credibility of the obtained results. While conducting this research, several relevant challenges were identified: a large number of hyperparameters, pseudo-randomness, complex experiment setups and models' sensitivity to changes in dataset. To address these and ensure reproducibility to the best extent possible, we follow the steps outlined in [34]:

The methodology section of the paper focuses on capturing clearly and concisely the steps that were followed through the study. We present every critical aspect and challenges encountered and support the solutions with suggestive visuals. Moreover, this work has been peer-reviewed two times. All together create the premise of reproducible experiments.

Notably, an extension to the written work is the open-sourced codebase[7]. This repository offers the immediate possibility of reproducing the work. Moreover, it allows for continuing the work immediately.

Furthermore, the complete dataset was acquired from the original authors of [7] while Twitter has granted the rights to fetch partially the dataset. We recommend those who would want to conduct their research with the complete dataset contact the authors of Moral Foundations Twitter Corpus paper[8] [7].

Lastly, computationally intensive operations were run on TU Delft HPC (High-Performance Computing). We present the environment and Python[9] libraries in Appendix A

7 Conclusion and Future Work

Understanding the moral drives behind sociocultural effervescence could unlock better policy-making, virtue-aligned AI and control of harmful movements. Such a moral profile, extracted from opinionated text, has made the subject of our work and is increasingly relevant to our society. Results, 72% Micro F1-score alongside 63% Macro F1-score (MFTC, multi-label classifier, 11 classes) are encouraging. We have tackled the problem from several technical angles: performance, generalizability and transferability. For each, we designed an adequate methodology.

Our study demonstrates that SBERT[10] and triplet loss [11] have a place in the study of moral rhetoric. With no related prior studies focusing on fine-tuning embeddings (meaningful vector representation of words and sentences), we developed a method to train them. Together with a simple Neural Network on top, they produce results that are above those achieved in [7] (53% averaged F1-score across foundations[10]) and in [8] (60% weighted average F1-score). Results vary across methodology, and we can conclude that we should train and test embeddings on the same domain of discourse for the most accurate results. Nonetheless, imbalanced datasets like Davidson provide little to no insight into the potential of our model with disappointing results of less than 10% F1-score. Once again, this suggests the urgent need for the expansion of MFTC with more adequate data. We also learned that values' recognisability varies across domain and that ALM and BLM datasets benefit the most from the fine-tuning methodologies. Sentence-BERT is a suitable architecture for this study, and triplet-loss can successfully embed morality in vectorial sentence representations. Once fine-tuned, our moral embeddings can serve more purposes such as question answering or clustering and are not limited to text classification.

7.1 Future Work

With only one rich dataset available (MFTC [7]), moral classifiers would benefit from more diverse and recent training data. For data collection, we identified a flexible methodology in [14] that filters tweets from a continuous stream (Twitter API) based on their hashtag (identifies the topic). However, after data is collected, its annotation can prove challenging, subjective and labour intensive. To partially mitigate these drawbacks, we recommend experimenting with semi-supervised annotating methods (e.g. [35]). Together,

[7] https://github.com/enricoliscio/nlp-for-values-CSE3000

[8] mdehghan@usc.edu

[9] https://www.python.org/

[10] F1-scores are reported for 5 binary classifiers, one per foundation. To compare, we averaged these scores.

they create a pipeline for extending the current state of Moral Foundations Twitter Corpus.

A proper understanding of why embeddings pre-trained for STS (Semantic Text Similarity) are suitable for moral classification would enable us to enhance the fine-tuning methodology. However, such a study falls within the subject of Explainability, "a critical problem in the field of Natural Language Processing" that focuses on "interpreting the outputs or the connections between inputs and outputs" [36] of a machine learning model. For this, we propose an experiment in which tweets' embeddings are clustered based on their cosine similarity. Then, for each resulting partition, we measure how similar are the foundations expressed in the tweets within it. A high overlap would confirm the hypothesis *semantically similar text expresses similar moral values*.

Lastly, the limited hyper-parameters exploration for fine-tuning and usage of only a set of SBERT pre-trained embeddings (out of many available[11]) leaves room for marginal performance increases. Therefore, the 72% Micro F1-Score is a low threshold for the true potential of the presented methodology.

8 Acknowledgements

I want to express gratitude to Professors Pradeep Murukannaiah and Enrico Liscio, my research supervisors, for their constructive feedback and great support in the development of this work. Thanks also extend to my colleagues Florentin Arsene, Ionut Constantinescu, Alin Dondera and Andrei Geadau, with whom I share part of the project's codebase[12].

References

[1] Jesse Graham, Jonathan Haidt, and Brian Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96:1029–46, 06 2009.

[2] Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. Moralization in social networks and the emergence of violence during protests. *Nature Human Behaviour*, 2(6):389–396, 2018.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020.

[5] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. Moralstrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction. *Knowledge-Based Systems*, 191:105184, Mar 2020.

[6] Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean Wojcik, and Peter Ditto. *Moral Foundations Theory*, volume 47, pages 55–130. 12 2013.

[7] Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

[8] Anne Fleur van Luenen. Recognising moral foundations in online extremist discourse : A cross-domain classification study. Master's thesis, Uppsala University, Department of Linguistics and Philology, 2020.

[9] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, Apr 2018.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015.

[12] Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. Linguistic knowledge and transferability of contextual representations, 2019.

[13] Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. Moralization in social networks and the emergence of violence during protests. *Nature human behaviour*, 2(6):389–396, 2018.

[14] Livia Teernstra, Peter van der Putten, Liesbeth Noordegraaf-Eelens, and Fons Verbeek. The morality machine: Tracking moral values in tweets. In Henrik Boström, Arno Knobbe, Carlos Soares, and Panagiotis Papapetrou, editors, *Advances in Intelligent Data Analysis XV*, pages 26–37, Cham, 2016. Springer International Publishing.

[15] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[16] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online, November 2020. Association for Computational Linguistics.

[17] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147, 2019.

[11] https://www.sbert.net/docs/pretrained_models.html

[12] https://github.com/enricoliscio/nlp-for-values-CSE3000

[18] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[19] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139–147, 2019.

[20] Nicholas Dingwall and Christopher Potts. Mittens: An extension of glove for learning domain-specialized representations, 2018.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[22] Sonali B Maind, Priyanka Wankar, et al. Research paper on basic of artificial neural network. *International Journal on Recent and Innovation Trends in Computing and Communication*, 2(1):96–100, 2014.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[24] Andrei Geadau. Performance analysis of the state-of-the-art nlp models for predicting moralvalues, 06 2021.

[25] Ionut Laurentiu Constantinescu. Evaluating interpretability of state-of-the-art nlp models for predicting moral values, 06 2021.

[26] Florentin-Ionut Arsene. Evaluating catastrophic forgetting of state-of-the-art nlp models for predicting moral values, 06 2021.

[27] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics.

[28] Charles Edwin Harris Jr., Michael Davis, Michael S. Pritchard, and Michael J. Rabins. Engineering ethics: What? why? how? and when? *Journal of Engineering Education*, 85(2):93–96, 1996.

[29] Michal Kosinski. Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports*, 11(1):100, 2021.

[30] Yilun Wang and M. Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114:246–257, 2018.

[31] Andrew Burt. Ethical Frameworks for AI Aren't Enough, 11 2020.

[32] Isabel Wagner and Eerke Boiten. Privacy risk assessment: From art to science, by metrics. In Joaquin Garcia-Alfaro, Jordi Herrera-Joancomartí, Giovanni Livraga, and Ruben Rios, editors, *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, pages 225–241, Cham, 2018. Springer International Publishing.

[33] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.

[34] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[35] Burr Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 1467–1478, USA, 2011. Association for Computational Linguistics.

[36] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy, July 2019. Association for Computational Linguistics.

A Environment

- GPU: GeForce RTX 2080 Ti
- Torch: 1.6.0
- TensorFlow: 2.2.0
- Sentence-Transformers: 1.1.1
- CUDA: 11.2
- cuDNN: 8.1.1.33

B Hyperparameters

Hyper-Parameters	Values
Epochs	[5, **10**, 15]
Activation	[**sigmoid** , relu]
Batch size	[16, **32**, 64]
Optimizer	[**Adam**]
Learning rate	[0.1, **0.01**, 0.05]
Threshold	[0.3, **0.4**, 0.5]

Table 9: Hyper-parameters, LSTM and Bi-LSTM

Hyper-Parameters	Values
Epochs	[3, 5, **10**]
Activation	[**sigmoid** , relu]
Batch size	[16, **32**, 64]
Optimizer	[**Adam**, AdamX]
Learning rate	[**0.0005**, 0.001]
Threshold	[0.3, **0.4**, 0.5]

Table 10: Hyper-parameters, NN on top of SBERT

Hyper-Parameters	Values
Epochs	[2, 5, **10**]
Batch size	[**16**, 32]
Loss function	[**BatchAllTripletLoss**]
Cost function	[**Euclidian**]
Margin	[**5**]

Table 11: Hyper-parameters, fine-tune SBERT

C Architectures

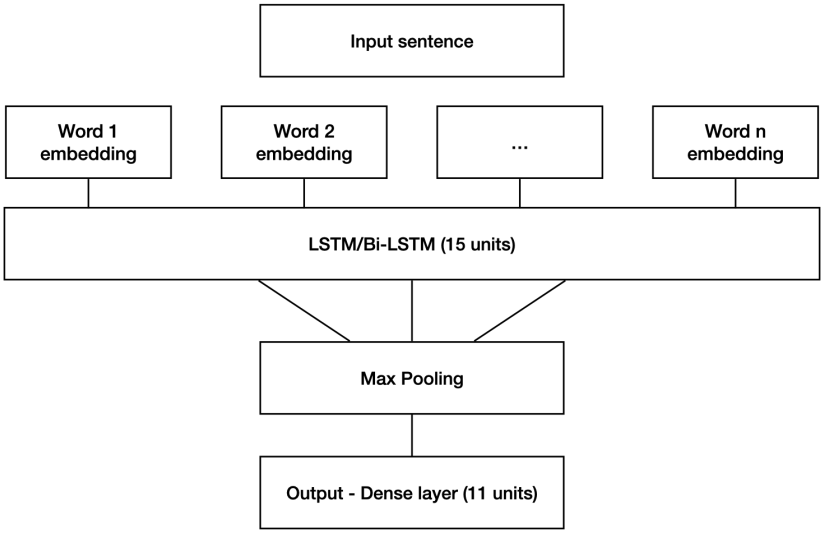

Figure 3: LSTM/Bi-LSTM with GloVe. Model architecture

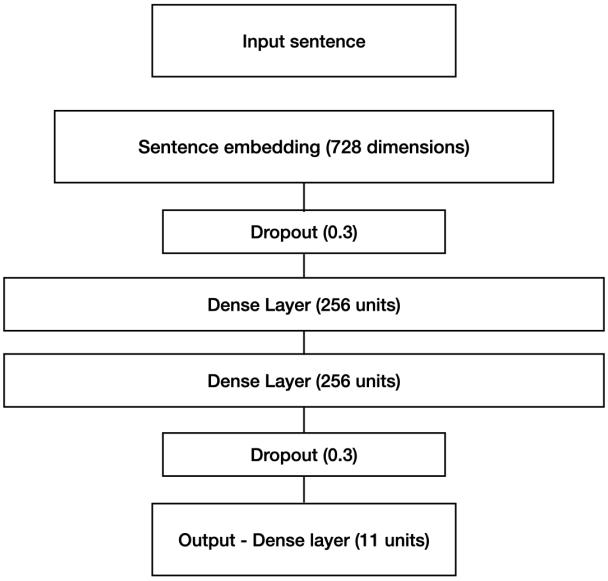

Figure 4: Simple NN with SBERT. Model architecture

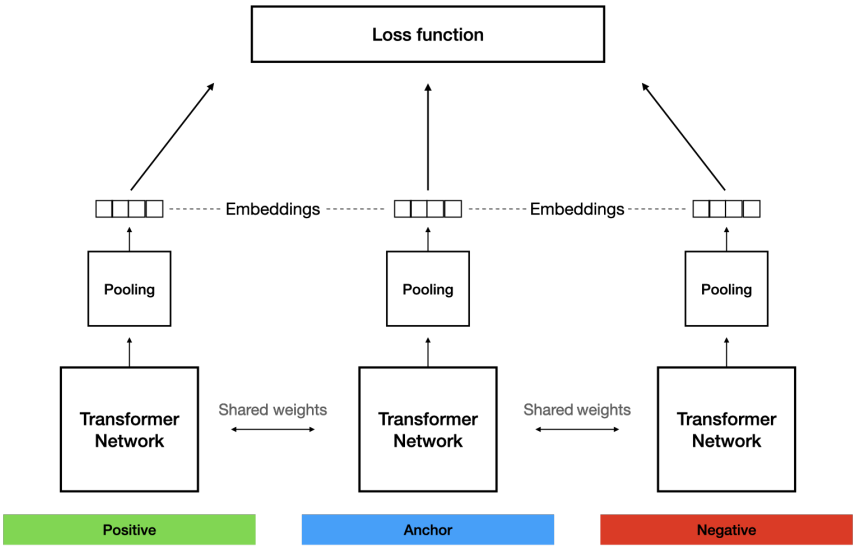

Figure 5: SBERT siamese triplet networks