Automatic Annotation of Dream Report's Emotional Content with Large Language Models

Lorenzo Bertolini **Valentina Elce** **Adriana Michalak**
Hanna-Sophia Widhoelzl **Giulio Bernardi** **Julie Weeds**
European Commission, Joint Research Centre (JRC), Ispra, Italy
MoMiLab Research Unit, IMT School for Advanced Studies, Lucca, Italy
Institute of Interdisciplinary Studies, University of Amsterdam, Amsterdam, The Netherlands
Department of Informatics, University of Sussex, Brighton, UK
lorenzo.bertolini@ec.europa.eu
{valentina.elce, adriana.michalak, giulio.bernardi}@imtlucca.it
hannasophia.widhoelzl@gmail.com
juliewe@sussex.ac.uk

Abstract

In psychology and neuroscience, dreams are extensively studied both as a model to understand the neural bases of consciousness and for their relationship with psycho-physical well-being. The study of dream content typically relies on the analysis of verbal reports provided upon awakening. This task is classically performed through manual scoring provided by trained annotators, at a great time expense. While a consistent body of work suggests that natural language processing (NLP) tools can support the automatic analysis of dream reports, proposed methods lacked the ability to reason over a report's full context and required extensive data pre-processing. Furthermore, in most cases, these methods were not validated against standard manual scoring approaches. In this work, we address these limitations by adopting large language models (LLMs) to study and replicate the manual annotation of dream reports, with a focus on reports' emotions. Our results show that a text classification method based on BERT can achieve high performance, is resistant to biases, and shows promising results on data from a clinical population. Overall, results indicate that LLMs and NLP could find multiple successful applications in the analysis of large dream datasets and may favour reproducibility and comparability of results across research.

1 Introduction

Dreams have fascinated humans since the dawn of time, and their scientific study in the last decades even increased attention and interest towards this peculiar phenomenon. Indeed, available evidence suggests that dreams may be related to psychophysical well-being, and may be involved in or represent a window on sleep-dependent processes affecting the consolidation and integration of new memories (Wamsley and Stickgold, 2011; Wamsley, 2014; Zadra and Stickgold, 2021). Moreover, given their nature of internally generated conscious experiences, dreams are regarded as a fundamental model to study and understand human consciousness (Nir and Tononi, 2010; Siclari et al., 2017). In spite of this, the mechanisms that lead to dream generation and development, and the possible functions of dreams still remain poorly understood to this day. Among the factors that limit and slow down research on dreams is the fact that the content of dreams is difficult to assess quantitatively and in a reproducible way (Elce et al., 2021).

Automating and standardising the scoring of dream reports' emotional dimensions is paramount for health and psychophysiological well-being as it can uncover valuable insights into an individual's mental states during sleep. As stated by the established continuity-hypothesis (Hall, 1953), elements in dream scenarios mirror someone's waking states and concerns (Brown and Donderi, 1986; Pesant and Zadra, 2006; Gilchrist et al., 2007; Blagrove et al., 2004). Nightmares have a particular potential to disrupt everyday life as they are linked to high levels of psychological distress, self-harm, and suicidal tendencies (Andrews and Hanna, 2020). Their frequency may serve as a promising early indicator of psychiatric and sleep disorders (Thompson et al., 2015; Kobayashi et al., 2008). In line with this, dream content was reported to change in several pathological conditions, including for instance eating disorders and depression (Skancke et al., 2014). Typically, the assessment of dream content

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 92–107
March 21, 2024 ©2024 Association for Computational Linguistics

— including the presence of specific emotions — is performed manually, by trained annotators, by applying particular scales or scoring systems. While multiple scoring approaches exist to annotate and analyse dream reports, such as the scale by Hauri and colleagues (Hauri, 1975) or the rating system developed by Schredl (Schredl, 2010), the Hall and Van de Castle (HVDC) coding system (Hall and Van De Castle, 1966) remains the most popular and widely adopted (McNamara et al., 2019; Fogli et al., 2020).

A growing body of evidence has shown that NLP methods can support the automatic analysis of dream reports. So far, efforts have mainly focused on investigating different implicit structures, such as speech or syntactic graphs (Mota et al., 2014; Martin et al., 2020), and/or analysing the semantic content of dream experiences (Sanz et al., 2018; Fogli et al., 2020; Zheng and Schweickert, 2021) (see Elce et al. (2021) for an extensive review). Of more relevance for this work are those studies that focused on dream reports' semantic content using dictionary-based linguistic analysis (Bulkeley and Graves, 2018; Mallett et al., 2021; Zheng and Schweickert, 2021) and distributional semantic models (Razavi et al., 2013; Altszyler et al., 2017; Sanz et al., 2018). While notably different, both approaches cannot fully and coherently manipulate a report's full content and context. This shared limitation is of great relevance, as the correct identification of an emotional state may rely on complex constructions and more implicit information, as well as a combination of the two. In recent years, these and similar issues were largely overcome by pre-trained large language models (LLMs) based on transformer architectures (Vaswani et al., 2017). Over the last years, LLMs pre-trained on self-supervised tasks like masked language models have shown strong performance on down-stream tasks like sentiment analysis (Raffel et al., 2020), text summarisation (Kedia et al., 2021), question answering (Lan et al., 2020), and machine translation (Conneau et al., 2020).

Given their success, in this work, we propose to address the issues identified with existing approaches to automatically analyse dream reports analysis by leveraging pre-trained LLMs. Specifically, we investigate whether and how LLMs can support the detection and analysis of emotions expressed in dream reports, as defined in accordance with the HVDC coding framework. More specifically, we study the ability of a bespoke multi-label text classifier, based on a pre-trained LLM tuned using dream reports previously scored by expert annotators, and propose a set of experiments and analyses to test the robustness of this solution to different potential biases in the dataset and out-of-distribution applications.

To the best of our knowledge, our work represents the first attempt to analyse and reproduce gold-standard HVDC annotations of dream reports with LLMs, and makes two main contributions. First, we show how, despite the limited amount of training data, a fully-supervised approach based on multi-label text classification yields good and stable performance. Two, we provide follow-up experiments and analysis showing how the strategies learned by the model are robust with respect to out-of-distribution data, as well as biases and spurious correlations present in the dataset.

2 Related Work

As summarised by Elce et al. (2021), a growing body of research is adopting NLP methods to automatically analyse dream reports. Yet, while emotions represent a fundamental component of oneiric experiences, only a fraction of published studies based on NLP methods have explicitly focused on the emotional aspects of dream reports (Nadeau et al., 2006; Amini et al., 2011; Razavi et al., 2013; Frantova and Bergler, 2009; McNamara et al., 2019; Fogli et al., 2020; Yu, 2022). Moreover, most of these investigations did not include a direct nor transparent comparison with widely adopted report annotation approaches such as the HVDC coding system. In terms of implementation, adopted NLP methods include three main solutions: graph-based approaches (Mota et al., 2014; Martin et al., 2020; Fogli et al., 2020), dictionary-based linguistic analysis (Bulkeley and Graves, 2018; Mallett et al., 2021; Zheng and Schweickert, 2021), and distributional semantic models (Razavi et al., 2013; Altszyler et al., 2017; Sanz et al., 2018).

Dictionary-based methods analyse data word by word, comparing each item to a dictionary file that is structured as a collection of words defining different semantic categories. An example could be the 'positive emotion' category, containing words such as *"joy"*, *"happiness"*, and *"smiling"*. Approaches based on these methods (Bulkeley, 2014; Bulkeley and Graves, 2018; Mallett et al., 2021; Zheng and Schweickert, 2021; Yu, 2022) are mainly used to determine the relative frequency of references to specific content words, and can hence be inherently misleading, as they generally cannot interpret contextual information and syntactic structures. Syntax-like structures are used by graph-based approaches, which cannot access semantics but have successfully been adopted to classify populations of participants (e.g., healthy and psychotic subjects)(Mota et al., 2014; Martin et al., 2020; Fogli et al., 2020). An exception is the work by Fogli et al. (2020), which proposed a solution based on a combination of dictionary and graph models, able to extract information about the content of dream reports, including their emotions. However, the evaluation was reframed in binary (*"positive"*, vs *"negative"*) terms.

Solutions based on distributional semantics (Nadeau et al., 2006; Razavi et al., 2013; Altszyler et al., 2017; Sanz et al., 2018; McNamara et al., 2019) were largely based on word-level representation obtained using models like `word2vec` (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). In such cases, encodings of full reports were generated by adding or averaging word embeddings, losing access to syntactic structure and more contextual understanding (Klafka and Ettinger, 2020).

A niche of previous work applied NLP and machine learning methods specifically to assess emotional aspects of dream reports (Amini et al., 2011; Razavi et al., 2013; McNamara et al., 2019; Yu, 2022). Amini et al. (2011) added word associations to improve the performance of a machine learning model that was trained to automatically score dream reports' emotional tone based on human judgements, resulting in increased machine-human agreement. Yet, relying on predefined word associations may oversimplify the nuances of emotional content unique to dreamers. Razavi et al. (2013) combined ad-hoc classifiers with a distributional approach to detect potential shifts in sentiment within each report. Evaluated reports were extracted from DreamBank, a large public database, and were (re-)scored by the authors using a four-level emotional rating system (*"very-negative"* to *"very-positive"*). Despite relying on a mixture of local (word-to-word) and general (sentence-to-sentence) occurrences, the adopted approach strongly relies on extensive data pre-processing, as well as composition by averaging, hence losing access to structural and deeper semantic information. McNamara et al. (2019) used a pre-trained agent to detect recurrent themes in a series of reports and found a partial match in the retrieved themes with aspects of the HVDC system, as well as significant differences in how themes occurred in male vs. female dreamers. The distribution of these themes was then used to assess the "mood" of each report. Yu (2022) combined a dictionary-based method with support vector machines (Cortes and Vapnik, 1995) to asses the general sentiment of dream reports in multiple languages.

Overall, the described studies present two main differences from our work. First, the models lacked access to the global context of each report. In our work, we use pre-trained large language models (LLMs) to encode dream reports and thus allow a model to have full access to the context. Second, the annotations and evaluations of emotional states were not directly compared against widely used coding systems such as the HVDC. That is, whenever human annotations were considered for the evaluation of a system, scores or labels were largely (and generally non-transparently) re-framed to be comparable with the output produced by the system of choice. In this work, we propose a solution to *adapt* a model to produce interpretable labels, that can be directly compared against human-produced HVDC annotations. Furthermore, evaluations will take into account the fact that labels could be associated with different characters, thus further highlighting the possible value of our approach as fully automatic and reliable support for manual annotations in dream research.

3 Dataset

For our experiments, we use a subset of reports extracted from the DreamBank database[1] (Domhoff and Schneider, 2008), pre-annotated according to the Hall and Van De Castle (HVDC) coding system (Hall and Van De Castle, 1966). `DreamBank.net` consists of a collection of over 20K dream reports gathered from different sources and organised in series, either provided by single individuals or groups of people, such as college students, teenagers, and blind adults. While `DreamBank.net` can be freely explored, the reports and the HVDC scores adopted in the current work are made available upon direct request to the researchers who maintain the DreamBank website. Among the approximately 1.8K labelled dream

[1] `https://www.dreambank.net`

reports, all in the English language, 922 contained at least one emotion associated with either the dreamer or another character. Considering that no guidelines or metadata are available to demonstrate that the absence of emotion labels reflects the *actual* absence of emotions in a report, we focus our experiments on those reports containing at least one emotion (n=922). The dataset is further divided into six series: `Bea 1: a high school student` (n=171/99; total number of reports/reports including at least one emotion), `Ed: dreams of his late wife` (n=143/108), `Emma: 48 years of dreams` (n=300/81), `Hall/VdC Norms: Female` (n=491/280), `Hall/VdC Norms: Male` (n=500/203), `Barb Sanders: baseline` (n=250/151).

The HVDC coding system examines ten categories of elements appearing in dream reports (characters, interactions, emotions, activities, striving, (mis)fortunes, settings and objects, descriptive elements, food and eating, and elements from the past). Within this study, we focused only on the annotation of the emotions feature. In the HVDC coding system, emotions are divided into 5 classes, that are anger (AN), sadness (SD), apprehension (AP), confusion (CO), and happiness (HA). Emotions might be assigned either to the dreamer or to other dream characters. We analyse both the emotions scored as experienced by the dreamer (*Dreamer Emotions*) and the overall occurrence of emotions in the dreams regardless of the dream characters they are associated with (*General Emotions*).

4 Multi-Label Text Classification

A set of preliminary experiments (see Appendix B) showed that an off-the-shelf sentiment analysis LLM cannot coherently solve the task when framed, similarly to previous work, as a binary `POSITIVE` - `NEGATIVE` classification. Hence, we investigate whether the human annotation of dream reports can be reproduced with supervision, via a bespoke text classification model, trained on gold-standard HVDC labelled data. Contrary to all previous work, that reframed HVDC labels to fit binary classification and classifiers, we perform a fine-grained classification aimed at determining the presence (1) or absence (0) of each HVDC emotion (i.e., *anger* (AN), *sadness* (SD), *apprehension* (AP), *confusion* (CO), *happiness* (HA)), regardless of the number of times they appear in a given report. Moreover, we experiment with both the sets of emotions described in Section 2: Dreamer Emotions and General Emotions.

4.1 Experimental Setup

Formally, we define the task as a multi-label classification, where a model is trained to simultaneously and independently predict *if* each of the emotions that were identified by expert annotators appear in each report. To solve the task, we designed a bespoke solution, where the LLM of choice is integrated into a three-component architecture, summarised in Figure 1. The first component is a pre-trained `BERT-large-cased` encoder[2], used to obtain the encoding of each report by extracting the final layer's `[CLS]` vector. Encodings are then fed to a dropout layer (with $p = .3$) and a linear layer, reducing the number of the dimensions to the number of desired classes, corresponding to the five HVDC emotions. The described architecture is then fully fine-tuned end-to-end with a binary cross-entropy loss, with the addition of a sigmoid function between the loss and the linear layer, and adopting a K-fold cross-validation procedure (with K=5). At each fold, the dataset is randomly split, 80% for training and 20% for testing, and the architecture is trained for 10 epochs, using dream reports as input and the presence of HVDC emotions as the output to predict. While the previous work evaluated a model on the HVDC annotation indirectly (e.g, by arbitrarily devising the five HVDC emotions into 2 classes) we evaluate the model *directly* on the HVDC's gold-standard annotation framework, by training and testing the model to simultaneously and independently guess if each of the five HVDC emotions was defined as present by the expert annotators (see Figure 1). Similarly to previous work investigating the presence of emotions in dream reports, we adopted precision, recall, and F1 as evaluation metrics (Fogli et al., 2020). The code is available here[3].

4.2 Results

Table 1 summarises the scores, averaged across the folds (± standard deviations) obtained by the architecture for Dreamer and General emotions. The overall F1 scores show a strong and generally stable performance. The minimal difference between

[2] To optimise the computational performance, we set the maximum length of the encoder to 512, losing full access to only 6 reports, accounting for less than the 0.005% of the whole dataset. See Appendix C.1 for more details.

[3] `https://github.com/lorenzoscottb/Dream_Reports_Annotation`

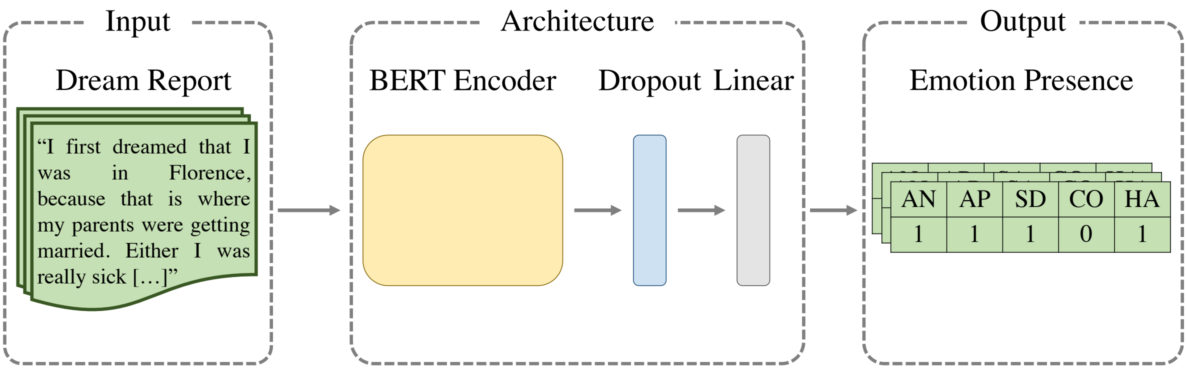
Input
Dream Report
"I first dreamed that I was in Florence, because that is where my parents were getting married. Either I was really sick [...]"
Architecture
BERT Encoder
Dropout
Linear
Output
Emotion Presence
AN
AP
SD
CO
HA
1
1
1
0
1

Figure 1: Schematic view of the adopted architecture and training procedure for the bespoke multi-label classification experiments. Given a set of dream reports from DreamBank as input, the architecture is trained end-to-end to predict which of the five emotions recognised by the Hall and Van de Castle (HVDC) system — **anger (AN)**, **apprehension (AP)**, **sadness (SD)**, **confusion (CO)**, and **happiness (HA)** — is present (1) or absent (0) in each report. The adopted architecture is constructed out of three components: a pre-trained LLM (in our case, a `BERT-large-cased model`), a dropout layer, and a linear layer.

	Precision		**Recall**		**F1**	
	Dreamer	General	Dreamer	General	Dreamer	General
Anger (AN)	86 ± 9	85 ± 7	89 ± 3	89 ± 4	87 ± 4	87 ± 5
Apprehension (AP)	86 ± 4	88 ± 7	88 ± 5	92 ± 3	87 ± 3	89 ± 3
Sadness (SD)	84 ± 10	84 ± 4	72 ± 15	77 ± 12	77 ± 11	80 ± 7
Confusion (CO)	90 ± 5	92 ± 2	76 ± 6	85 ± 5	82 ± 5	88 ± 3
Happiness (HA)	93 ± 5	86 ± 4	85 ± 6	88 ± 3	89 ± 5	87 ± 2
macro avg	88 ± 3	87 ± 3	82 ± 5	86 ± 2	85 ± 3	86 ± 2
micro avg	87 ± 3	87 ± 3	84 ± 4	87 ± 2	85 ± 3	87 ± 2
samples avg	88 ± 2	89 ± 2	87 ± 3	90 ± 2	86 ± 3	88 ± 2
weighted avg	88 ± 2	87 ± 4	84 ± 4	87 ± 2	85 ± 3	87 ± 2

Table 1: Bespoke muli-label classification results. Average scores (± standard deviation) of the 5-fold cross-validation text classification experiment. Dreamer and General columns refer to the Emotions used for training and testing. While under the General Emotions setting we made use of all emotions found by expert annotators in each report, the Dreamer Emotions refers to the subset of the General Emotions associated by the expert annotators solely to the dreamer.

macro and weighted F1 scores further suggests that the difference in support instances only has a marginal impact. Concerning single Emotion sets, performance tends to be higher and more stable for General than for Dreamer emotions. When trained and tested on General emotions, the models show a notable balance between precision and recall, despite a relatively higher variance across precision measures. On the other hand, models trained solely with Dreamer emotions present an overall higher precision than recall, with the latter being notably less stable. These patterns are likely explained by the low number of emotions-per-report associated with the Dreamer set, while the emotion distribution is more balanced in the General set. Models trained solely with the Dreamer set are hence less prone to produce False-Positive errors but produce a higher amount of False-Negative errors. Since the General emotion set is overall more balanced, the models' performance is higher and more stable across precision and recall.

Concerning single emotions, it is more difficult to identify a shared pattern, with the notable exception of *sadness* (SD). Under both sets, models appear to struggle at classifying such an emotion, which in both cases produces the highest variance, an observation that might be partially explained by sadness being the least frequent emotion.

Our results indicate that the model can successfully learn to simultaneously classify a dream report with respect to references to the different emotions of the HVDC coding system. However, the achieved performance level might be mediated, at least in part, by specific series of DreamBank. It is in fact possible that different emotions are distributed in a particular and unique way in each series. If so, the model could learn series-specific distributions, and, after implicitly recognising a specific series in a given report, simply reproduce these distributions at test time. For example, if a series like `Ed` contained a large number of reports labelled both with *sadness* and *apprehension*, the model could implicitly learn to identify `Ed`'s reports from such series via recurrent cues to unrelated information (such as characters or places) and, at test time, use these cues to automatically annotate those reports with *sadness* and *apprehension*.

4.2.1 Ablation

To understand whether the performance of the trained model is affected by this heuristic behaviour — that is, learning series-specific emotion distributions — we conduct a follow-up ablation experiment. Using the same architecture, hyperparameters, and training setup, instead of randomly splitting five times the whole dataset into an 80-20% train-test split, we here use one whole series of the dataset as the test set and the remaining series as the training set. With this approach, test series are never seen by the model during training, making it impossible for the model to rely on series-specific associations for solving the task. For this experiment, we focus solely on the General Emotions set, found to be the best-performing and more stable set. Moreover, we focus the analysis on the F1 scores as the performance metric of choice.

Figure 2 summarises the results of the ablation experiment. The x-axis shows the F1 weighted average scores obtained for each series (y-axis) when such a series is held out from training and used as the test set. In order to facilitate comparison with the previous experiment's results, the dashed grey line indicates the F1 weighted average obtained in the K-fold experiment (i.e., 87 ± 2, see Table 1). The results indicate that when all the instances of a series are removed from the training data, the test performance of the model remains relatively high and stable. Moreover, as shown in Figure 3 this remains true for all of the HVDC scored emotions.

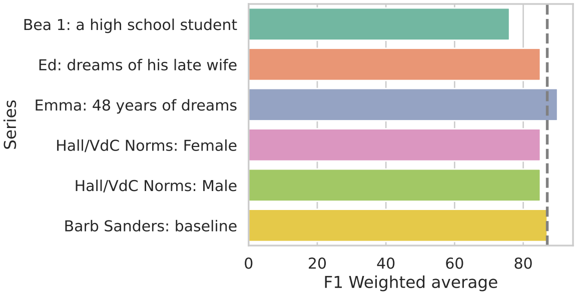

Figure 2: Ablation experiment results. F1 Weighted average scores obtained by the model when each Dream Bank's Series is held out of training and used as a test set. The dashed vertical line reports the average F1 Weighted average obtained in the main experiment (see Table 1).

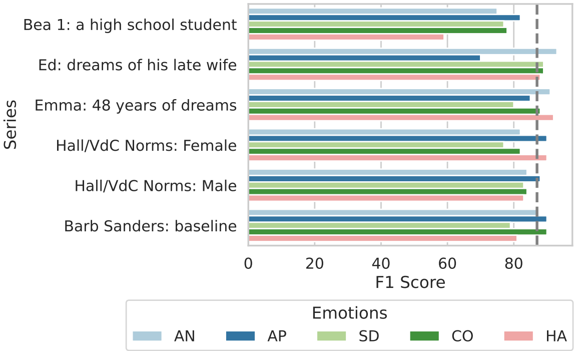

Figure 3: Ablation experiment, emotion by Series analysis. The diagram further breaks down the results of Figure 2 by single emotion for each Series held out from training. Once again, the vertical dotted line refers to average scores in the main experiment (see Table 1).

The `Bea 1` series, appears to represent the only notable exception to the above observations. Indeed, this series shows the greatest deviation from the original results, with an F1 weighted average of 77, compared to the previously obtained average of 87 (± 2). As shown in Figure 3, which breaks down the results of the ablation experiment presented in Figure 2 by single emotions, this was largely due to a problematic classification of *happiness* (HA) in this particular series. However, with the exception of a slightly lower *sadness*, emotions don't seem to significantly deviate from the K-fold experiment results, as summarised by Figure 4.

These results support two main conclusions. First, the proposed architecture, based on a pre-trained LLM, can learn efficient classification strategies for dream reports' emotional content (as defined based on the HVDC coding system). Second, the learned model does not rely on simple heuristics based on series-dependent cues and dis-

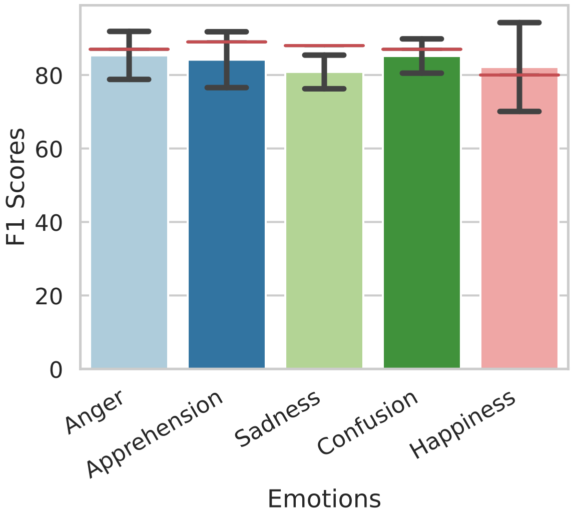

Figure 4: Ablation experiment, single emotion analysis. Overall results (in F1 scores) for every single emotion obtained in the different Series for the ablation experiment (see Figure 3). Bars report standard deviation, while Red horizontal lines refer to average scores in the main experiment (see Table 1).

tributions. That said, ablation's results could have been influenced by yet another confound: the numbers of reports and emotion distributions. In other words, the performance of each combination of series-emotion (e.g., `Bea 1`-*happiness*) could be explained by the number of items provided at test time. To assess this possibility we perform a set of series-independent Spearman's correlations between the number of test items for each emotion and their respective results (i.e., the F1 scores). The results, summarised in Appendix C.3 found no connections between F1 scores and the number of test items.

4.2.2 Out of distribution PTSD data

So far, results suggest that our solution could provide a valuable resource to annotate data even from out-of-distribution participants. However, annotated data contain reports solely from healthy individuals. Since dream reports can provide useful information on the mental state of an individual, it would be important to assess the robustness of the model to data from participants of different clinical populations. To test this, we adopt a series, not contained in the annotated version of the dataset, containing reports from a Vietnam war veteran with a diagnosed post-traumatic stress disorder (PTSD), who had frequent negative dreams and nightmares. While we do not have an *actual* emotion distribution for such a series, we can *assume* an expected one, with a strong predominance of negatively connotated emotions. We fine-tuned a model using the same data, architecture, hyper-parameters, and procedure thus far adopted, with the only addition of an early-stop mechanism and no K-fold.

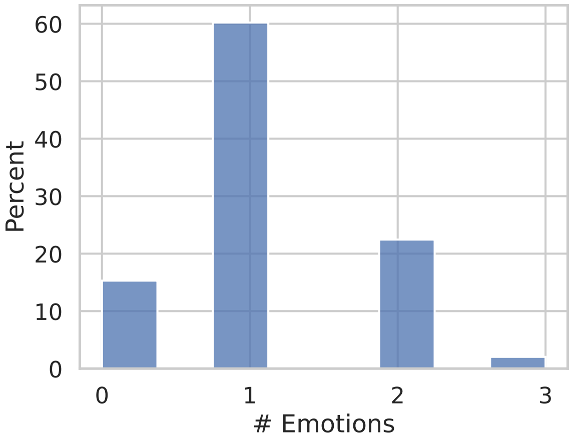

Figure 5: Number of predicted emotions per report. Distribution of the number of emotions-per-report predicted by the model for the `Veteran` series.

Out of the 98 dreams contained in the `Veteran` series, the model found at least one emotion in approximately 84% of them. As summarised by Figure 5, most of these reports were associated by the model with a single emotion, and approximately 20% of them were labelled with two emotions. As expected, the vast majority of these reports contain negatively connotated emotions, as seen in Figure 6. *Apprehension* is by far the most observed negative emotion, appearing in more than half of the reports. Moreover, Figure 6 strongly suggests that the emotion distribution proposed by the model for the `Veteran` series is not simply a transposition of the one observed by the model during training. This further suggests that the model has successfully learned reliable and generalised classification strategies, and it does not simply reproduce an observed distribution from the training data.

The model also annotated a minority of reports – circa 19% – with *happiness*. A manual inspection did identify some errors but also found multiple instances where the model's annotation (i.e., including happiness as an emotion expressed within the report) seemed justified. For example, in one of these reports, after describing a very violent war scenario, the Veteran adds that he felt *"a feeling of complete freedom. In very high spirits Jim L. and I go to a supermarket and buy food. I am aware that I don't wear my steel helmet."*. In another case,

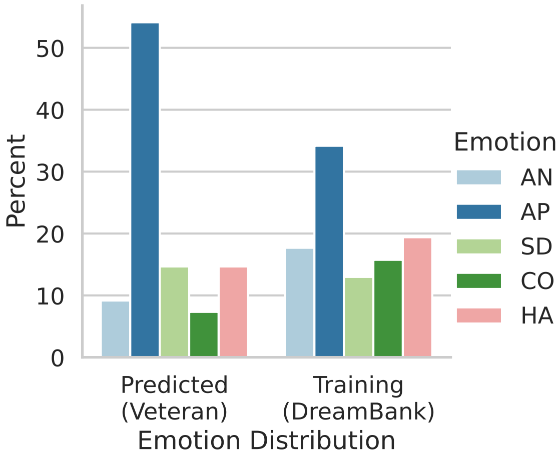

Figure 6: `Veteran` and DreamBank emotion distributions. Comparison of the emotion distribution predicted by the model for the (unlabelled) `Veteran` series, and the general emotion distribution in the DreamBank dataset, used to train the model.

a dream resembling a nightmare (two dolls have come to life) is narrated in a normal and friendly manner, as clear from the passages *"I speak to the male doll and the female doll and feel happy. I have made two friends."*. In other cases, the negative context of the dream is notably less dominant, and the report simply describes a series of social encounters and interactions. We also found scenarios clearly triggered by strong cue words and context. For instance, in one report the Veteran is in a rehab clinic, surrounded by other veterans, and children's paintings, and adds that they are *"[...] colourful, lively, happy. There is no sense of war"*; in another, he describes a romantic encounter – *"We are happy and young. She puts her arm around my shoulder. " I like you, " she says. " I really like you.""*.

5 Discussion

In the field of dream research, the assessment of a report's emotional content is typically based on time-consuming, annotator-dependent procedures. Throughout the years, only a few studies employed automated approaches based on NLP techniques, including dictionary-based and distributional semantics methods. However, these approaches have very limited access to the syntax and semantics of a report's content, and may thus fail to correctly and fully capture emotions described in dream reports. In this work, we tested whether a transformer-based large language model (LLM) could be used to overcome such limitations and reproduce human-based scoring. Specifically, we trained a model end-to-end, using pre-annotated data to predict *if* and *which* emotions were present or absent in a given dream report. The obtained results showed that the model was able to learn reliable and stable classification rules. Follow-up experiments further confirmed two important aspects of our solution. First, via an ablation experiment, we showed that the ability of the model to solve the task is only marginally affected by differences between distinct subsets of the training data. Second, such generalisation holds also for instances that significantly deviate from the training data, as shown by the experiment with reports from a PTSD patient.

Our findings suggest that what is more likely to impact the model performance is the vocabulary used to describe specific emotions across different series. Indeed, variability in the used vocabulary may be explained by the fact that the series included in the present work were collected from different individuals or groups of individuals, with relevant differences in demographic, psychological, and behavioural characteristics. Should this be the case, it would be yet another reason to support the use of tools that are able to reason over the full content of a report, and have access to a large and dynamic vocabulary, already have significant information about a large set of lexemes, and can be easily adaptable to new words and languages. Current pre-trained transformer-based LLMs satisfy all these requirements. Given the current state of NLP resources, our proposed architecture can be easily adapted to be used with LLMs pre-trained on different languages or tasks. Moreover, trained models like ours are fully open-source and can be easily adopted by researchers and practitioners in their pipeline, making these results and the framework extremely replicable and widely standardised.

6 Conclusion

In this work, we tested the feasibility of using transformer-based large language models (LLM) to annotate dream reports with respect to emotions expressed in a given report. Our results show that our LLM-based solution using multi-label classification yields a strong performance, which was found to be robust to correlational biases and out-of-distribution data. Such approaches have the potential to significantly accelerate research investigating the origin, meaning, and functions of dreams, and might present a valuable and efficient support or alternative to human-based procedures involving the analysis of large datasets, ensuring at the same time reproducibility of the obtained results through the sharing of adopted models.

7 Limitations

This study presents three main limitations. First, while DreamBank does contain reports in multiple languages, the HVDC annotations were available only for reports in English. Thus, the generalizability of our model and observations to other languages remains to be determined. Second, while the dataset under consideration was relatively large with respect to studies in the field of dream and sleep research, it is instead relatively small for a machine-learning investigation, especially for the use of supervised methods. Third, the success of the model can be interpreted only to the emotion feature of the HVDC framework. While we provide extensive experiments and evidence supporting the generalisation ability of our model, these are limited specifically to emotion-based annotations. Future work will have to assess the feasibility of our solution to other aspects and features of the HVDC framework, which might require a different approach rather than text classification systems, such as text-to-text generation models.

8 Ethical and Broader Impact Considerations

It is important to acknowledge that, while we have stressed the link between dream reports and mental health, our study and proposed model should only be considered from a basic research perspective. Our procedure and trained model have no diagnostic valence, and should only be considered as a tool to support the annotation of a (large) set of dream reports only from an experimental and hypothesis-building perspective, always keeping in mind the inevitable limitations that come from adopting a machine in the annotation process.

Acknowledgements

This research was supported by the EU Horizon 2020 project HumanE-AI (grant no. 952026), and a BIAL Foundation Grant (grant no. 091/2020). We thank Ian Morgan Leo Pennock, Giacomo Handjaras, and the anonymous reviewers for their comments on the work.

References

Edgar Altszyler, Sidarta Ribeiro, Mariano Sigman, and Diego Fernández Slezak. 2017. The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. *Consciousness and Cognition*, 56:178–187.

Reza Amini, Catherine Sabourin, and Joseph De Koninck. 2011. Word associations contribute to machine learning in automatic scoring of degree of emotional tones in dream reports. *Consciousness and cognition*, 20(4):1570–1576.

Sophie Andrews and Paul Hanna. 2020. Investigating the psychological mechanisms underlying the relationship between nightmares, suicide and self-harm. *Sleep medicine reviews*, 54:101352.

Mark Blagrove, Laura Farmer, and Elvira Williams. 2004. The relationship of nightmare frequency and nightmare distress to well-being. *Journal of Sleep Research*, 13(2):129–136.

Ronald J Brown and Don C Donderi. 1986. Dream content and self-reported well-being among recurrent dreamers, past-recurrent dreamers, and nonrecurrent dreamers. *Journal of Personality and Social Psychology*, 50(3):612.

Kelly Bulkeley. 2014. Digital dream analysis: A revised method. *Consciousness and cognition*, 29:159–170.

Kelly Bulkeley and Mark Graves. 2018. Using the LIWC program to study dreams. *Dreaming*, 28(1):43–58.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

G. William Domhoff and Adam Schneider. 2008. Studying dream content using the archive and search engine on DreamBank.net. *Consciousness and Cognition*, 17(4):1238–1247.

Valentina Elce, Giacomo Handjaras, and Giulio Bernardi. 2021. The language of dreams: Application of linguistics-based approaches for the automated analysis of dream experiences. *Clocks & Sleep*, 3(3):495–514.

Alessandro Fogli, Luca Maria Aiello, and Daniele Quercia. 2020. Our dreams, our selves: automatic analysis of dream reports. *Royal Society Open Science*, 7(8):192080.

Elena Frantova and Sabine Bergler. 2009. Automatic emotion annotation of dream diaries. In *Proceedings of the analyzing social media to represent collective knowledge workshop at K-CAP 2009, The fifth international conference on knowledge capture*. Citeseer.

Sue Gilchrist, John Davidson, and Jane Shakespeare-Finch. 2007. Dream emotions, waking emotions, personality characteristics and well-being–a positive psychology approach. *Dreaming*, 17(3):172.

Calvin S Hall. 1953. The meaning of dreams.

Calvin S. Hall and Robert L. Van De Castle. 1966. *The Content Analysis of Dreams.* Appleton-Century-Crofts.

P Hauri. 1975. Categorization of sleep mental activity for psychophysiological studies. *The experimental study of sleep: Methodological problems*, pages 271–281.

Akhil Kedia, Sai Chetan Chinthakindi, and Wonho Ryu. 2021. Beyond reptile: Meta-learned dot-product maximization between gradients for improved single-task regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 407–420, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Ihori Kobayashi, Eve M Sledjeski, Eileen Spoonster, William F Fallon Jr, and Douglas L Delahanty. 2008. Effects of early nightmares on the development of sleep disturbances in motor vehicle accident victims. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, 21(6):548–555.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Remington Mallett, Claudia Picard-Deland, Wilfred Pigeon, Madeline Wary, Alam Grewal, Mark Blagrove, and Michelle Carr. 2021. The relationship between dreams and subsequent morning mood using self-reports and text analysis. *Affective Science*, 3(2):400–405.

Joshua M. Martin, Danyal Wainstein Andriano, Natalia B. Mota, Sergio A. Mota-Rolim, John Fontenele Araújo, Mark Solms, and Sidarta Ribeiro. 2020. Structural differences between REM and non-REM dream reports assessed by graph analysis. *PLOS ONE*, 15(7):e0228903.

Patrick McNamara, Kelly Duffy-Deno, Tom Marsh, and Thomas Marsh. 2019. Dream content analysis using artificial intelligence. *International Journal of Dream Research*, Vol 12:No 1 (April 2019).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Natália B. Mota, Raimundo Furtado, Pedro P. C. Maia, Mauro Copelli, and Sidarta Ribeiro. 2014. Graph analysis of dream reports is especially informative about psychosis. *Scientific Reports*, 4(1).

David Nadeau, Catherine Sabourin, Joseph De Koninck, Stan Matwin, and Peter D. Turney. 2006. Automatic dream sentiment analysis. Proc. of the Workshop on Computational Aesthetics at the Twenty-First National Conf. on Artificial Intelligence.

Yuval Nir and Giulio Tononi. 2010. Dreaming and the brain: from phenomenology to neurophysiology. *Trends in Cognitive Sciences*, 14(2):88–100.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nicholas Pesant and Antonio Zadra. 2006. Dream content and psychological well-being: A longitudinal study of the continuity hypothesis. *Journal of clinical psychology*, 62(1):111–121.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Amir H. Razavi, Stan Matwin, Joseph De Koninck, and Ray Reza Amini. 2013. Dream sentiment analysis using second order soft co-occurrences (SOSCO) and time course representations. *Journal of Intelligent Information Systems*.

Camila Sanz, Federico Zamberlan, Earth Erowid, Fire Erowid, and Enzo Tagliazucchi. 2018. The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports. *Frontiers in Neuroscience*, 12.

Michael Schredl. 2010. Dream content analysis: Basic principles.

Francesca Siclari, Benjamin Baird, Lampros Perogamvros, Giulio Bernardi, Joshua J LaRocque, Brady Riedner, Melanie Boly, Bradley R Postle, and Giulio Tononi. 2017. The neural correlates of dreaming. *Nature Neuroscience*, 20(6):872–878.

Joacim F Skancke, Ingrid Holsen, and Michael Schredl. 2014. Continuity between waking life and dreams of psychiatric patients: a review and discussion of the implications for dream research. *International Journal of Dream Research*.

A Thompson, Suzet Tanya Lereya, G Lewis, Stanley Zammit, Helen L Fisher, and Dieter Wolke. 2015. Childhood sleep disturbance and risk of psychotic experiences at 18: Uk birth cohort. *The British Journal of Psychiatry*, 207(1):23–29.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Erin J. Wamsley. 2014. Dreaming and offline memory consolidation. *Current Neurology and Neuroscience Reports*, 14(3).

Erin J. Wamsley and Robert Stickgold. 2011. Memory, sleep, and dreaming: Experiencing consolidation. *Sleep Medicine Clinics*, 6(1):97–108.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Calvin Kai-Ching Yu. 2022. Automated analysis of dream sentiment—the royal road to dream dynamics? *Dreaming*, 32(1):33–51.

Antonio Zadra and Robert Stickgold. 2021. *Dreaming and Offline Memory Consolidation*. WW Norton & Company.

Xiaofang Zheng and Richard Schweickert. 2021. Comparing hall van de castle coding and linguistic inquiry and word count using canonical correlation analysis. *Dreaming*, 31(3):207–224.

A DreamBank's Distributions

The section presents more details and analyses of DreamBank's statistics. Figure 7 shows the distribution of the HVDC emotions in DreamBank, divided between the different series of DreamBank. Figure 8 summarises how single DreamBank reports distribute with respect to the number of (General) emotions per report. As shown, the majority (circa 65%) of the 922 reports containing at least one emotion in fact contain only one emotion. Approximately 25% contains two emotions, while the rest can reach up to 9 emotions per report. When considering only Dreamer emotions, the percentage of reports with only one emotion reaches almost 75%, and the number of reports with more than two emotions drops to approximately 5% of the total (see Figure 9).

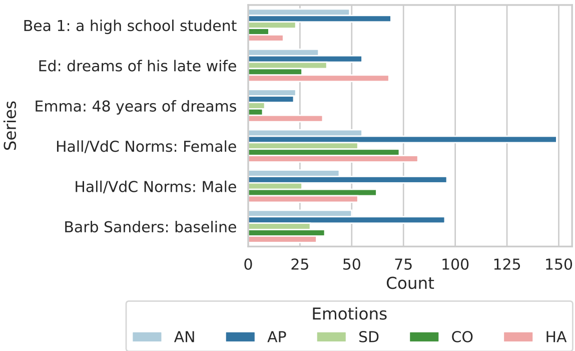

Figure 7: General emotion distribution across Dream Bank's Series.

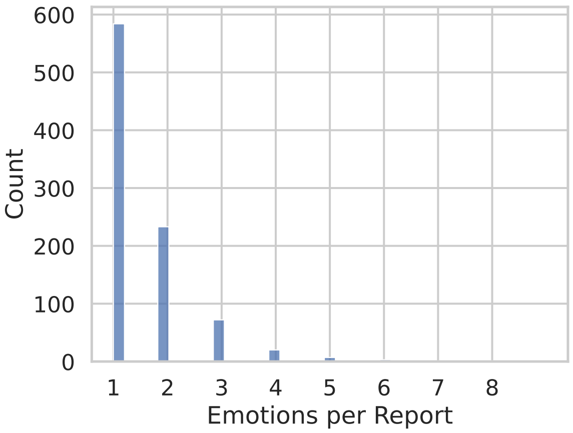

Figure 8: Number of emotion per report. Visualisation of how reports distribute with respect to the number of (General) emotions they have been labelled with.

B Off-the-Shelf Sentiment Analysis

We here discuss the results of a two-level preliminary experiment, where we investigated if an off-the-shelf model tuned to perform sentiment analysis (SA) could have been used to assess the emotional content of dream reports. Specifically, we proposed to test a 2-way `POSITIVE` vs. `NEGATIVE` classification, similar to previous work (McNamara et al., 2019; Yu, 2022). The experiment was run using the default SA setting of Hugging Face's (Wolf et al., 2020) `pipeline`, and had two levels. First, we investigated whether the general predictions of the model (i.e., the predicted labels and their scores) correlated with the

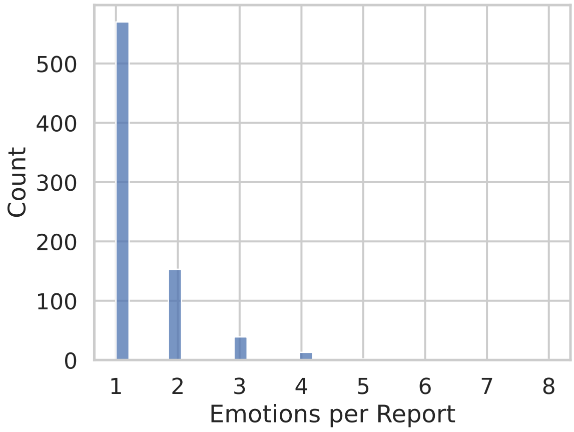

Figure 9: Number of Dreamer-only emotion per report. Visualisation of how reports distribute with respect to the number of (Dreamer-only) emotions they have been labelled with.

sentiment of individual dream reports. We defined the overall sentiment of each report as the sum of all references to emotions identified according to the HVDC coding system. A schematic summary of our approach is presented in Figure 10.

The second experiment focused on those reports containing a single emotion, and studied whether the predicted label (i.e., `POSITIVE` or `NEGATIVE`) matched the emotion found by the annotators.

B.1 Annotator Score

While 90% of the gathered dream reports do not present more than two emotions, some reports can contain a large variety of emotions — up to 9 in some rare cases (see Appendix A for more details). Hence, the main aim of the sentiment analysis investigation was to assess whether the model's predictions do reflect the overall *sentiment* of a report, defined according to the number of times specific positive or negative emotions appear in a report (regardless of the character who experienced them). Formally, given a dream report containing a list of Emotions E, such as the one in the example of Figure 10, and a scoring table S, mapping each HVDC emotion to a set having positive (E_+), negative (E_-), or neutral (E_0) valence, we computed the sentiment of a report (i.e., the *Annotator Score* (AN)) through the equation in 1

$$AN(E) = P(E_+) - P(E_-) \tag{1}$$

with

$$P(E_+) = \frac{|E_+|}{|E|} \tag{2}$$

and

$$P(E_-) = \frac{|E_-|}{|E|} \tag{3}$$

For this experiment, our scoring table S assigned *anger*, *apprehension*, *sadness* to the negative valence set (E_-), *happiness* to the positive valence (E_+) set, and *confusion* to the neutral set (E_0) (see Figure 10 for an example). Similarly, the *Model Score* of a report was defined as the difference between the probability associated with the `POSITIVE` and `NEGATIVE` labels. For instance, if the model predicts the probability distribution of the `POSITIVE` ($P(+)$) and `NEGATIVE` ($P(-)$) labels to be .4 and .6, respectively, then the *Model Score* for such a report would be –.2 (see Figure 10 for an example).

The model's performance was assessed by comparing the *Model Score* with the *Annotator Score* via Spearman's correlation coefficient (ρ).

B.1.1 Results

Figure 11 presents the results of the correlation analysis between scores produced by human annotators and the selected model, for the Dreamer and General Emotion sets. While the correlation with the General Emotions is marginally better, results are overall poor. Moreover, under both Dreamer and General Emotions, the performance was heavily influenced by different DreamBanks' series, as demonstrated by Figure 12. Interestingly, under both the Dreamer and the General Emotions, `Ed` and `Emma` seem to present the strongest correlation between human and model scores.

Figure 13 suggests that these results were likely due to the slightly different distributions produced by human annotators and the sentiment analysis model. Indeed, the predictions of the model (i.e., the *Model Scores*, x-axis) were strongly polarised. In other words, the model was consistently very confident in its decisions on which sentiment (`POSITIVE` or `NEGATIVE`)) was appearing in a given report. On the other hand, the *Annotator Scores* (y-axis) presented a cluster of instances around the value of 0. Interestingly, a considerable part of these reports contained two or three emotions (see Appendix B.3 for more details). Given the adopted method to compute *Annotator Scores* (see Eq. 1), such cluster presents a high number of instances annotated with a single positive emotion and a single negative emotion, or those two plus *confusion*. The fact that such a cluster of

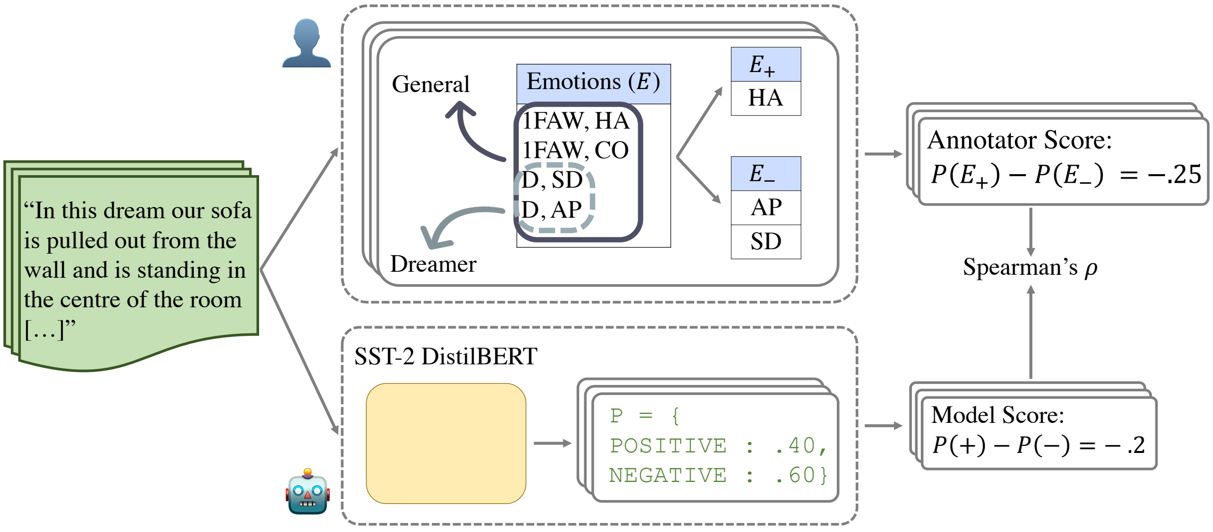

Figure 10: Proposed setup for the Annotator Score experiment (Section B.1).

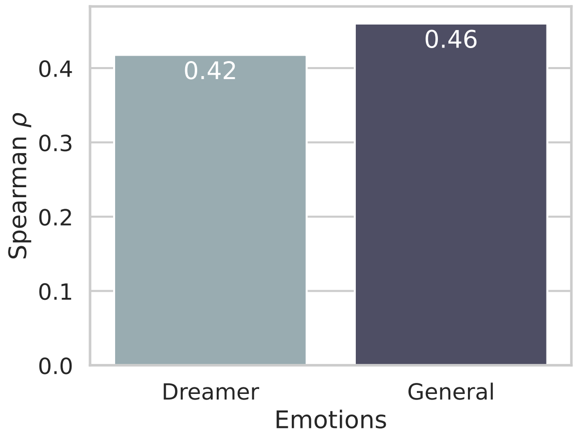

Figure 11: Report sentiment results. Correlations coefficients (in Spearman’s ρ) between the model predictions and dream report’s sentiment.

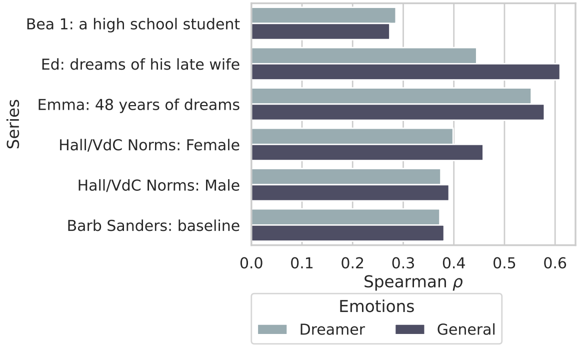

Figure 12: Report sentiment, collection analysis. Correlations coefficients (in Spearman’s ρ) between the model predictions and each report’s sentiment, divided by Dream Bank’s Series.

zero-valued *Annotator Scores* containing conflicting emotions did not have a clear match in the *Model Scores* distribution, suggests that the model might be picking on either the positive or negative emotion. Therefore, the scores of the model may not efficiently reflect the more general sentiment of the reports, but only encode the presence of a specific emotion type (positive or negative). The following experiment investigated this possibility, focusing on those reports only containing one emotion, and approaching the problem from a categorical perspective.

B.2 Single-Emotion

The first experiment showed how the selected model failed to correctly capture the distribution of human annotators’ scores, mainly due to very polarised predictions and might have simply reflected what type of emotion (positive or negative) is mainly present in a given report. Since the HVDC system also allows assigning a strictly positive or negative connotation to each emotion, we studied such a possibility by focusing solely on those reports that experts have annotated with one — and only one — of the five HVDC emotions: *anger*, *apprehension*, *confusion*, *sadness* and *happiness*. The goal was thus to understand if reports classified as `POSITIVE` or `NEGATIVE` by the model do contain an emotion that the HVDC scoring system also defined as positive or negative. Here, results are interpreted in terms of precision, recall and F1, with respect to the two prediction classes (`POSITIVE` and `NEGATIVE`).

	Precision		**Recall**		**F1**	
	Dreamer	General	Dreamer	General	Dreamer	General
`NEGATIVE`	92	91	83	82	87	86
`POSITIVE`	44	45	64	65	52	53
macro avg	68	68	73	73	70	70
weighted avg	83	82	79	78	81	80

Table 2: Single-emotion results. Per-class and average scores obtained when comparing model-predicted and human-generated labels for dream reports containing a single emotion. Here, the five HVDC emotions were collapsed into positive (i.e., *happiness*) and negative (i.e., *anger*, *apprehension*, *sadness* and *confusion*), and compared against the label predicted by the sentiment analysis model (i.e., `POSITIVE` or `NEGATIVE`).

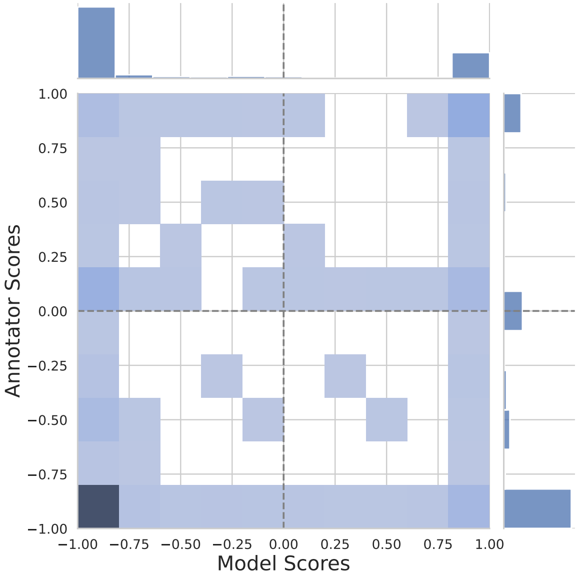

Figure 13: Annotator Score, predictions' analysis. Comparison of the Predicted Sentiment scores (x-axis) and Report Sentiment distribution (y-axis) for the General Emotion set. As seen, while the model's predicted scores are strongly polarised, annotators' scores, computed via Eq. 1, are more smoothly and evenly distributed.

B.2.1 Results

Table 2 summarises the results and clearly shows that, with respect to reports containing a single emotion, the predictions of the model matched the human-produced annotations only with respect to negative emotions, while showing poor results with respect to the `POSITIVE` class —– which only contains *happiness*. The model was however largely unstable with respect to the type of error it makes, as shown by the notable difference between precision and recall scores.

Figure 14 presents the same results of Table 2, divided by single HVDC emotion (x-axis) and series (diagrams), and shows how the model remained notably impacted by the different DreamBank's series. Of note, `Ed` and `Emma`, the two series that produced the best performance in the previous experiment, here showed the most balanced results across different HVDC emotions. Overall, these results strongly suggest that the selected model had fewer problems when classifying reports containing negative emotions than at detecting the presence of positive emotions.

B.3 Annotator vs. Model Scores Analysis

The section presents a more detailed analysis of the distributions of Model and Annotator Scores, with respect to the number (#) of emotions. As shown in Figure 15, the two peaks of the `Model Scores` distributions mainly contained reports classified by annotators as presenting a single emotion. However, the proportion of reports containing two emotions is notably higher in those reports classified by the model as being strongly `NEGATIVE`. Interestingly, with respect to the `Annotator Scores`, the proportion of reports with two emotions is concentrated in those reports with `Annotator Scores` of –2 (see Figure 16)

C Multi-Label Text Classification

C.1 Token distribution

Figure 17 summarises the distribution of tokens produced by the tokeniser of the selected pretrained LLM (i.e., `BERT-large-cased`), divided by DreamBank's series. As seen, only 6 reports, accounting for approximately 0.003% of the whole dataset, present more than 510 content tokens.

C.2 Supervised Learning Hyper-Parameters

Table 3 collects the hyper-parameters used to tune the bespoke classifiers from Section 4. The same parameters were used throughout the whole work.

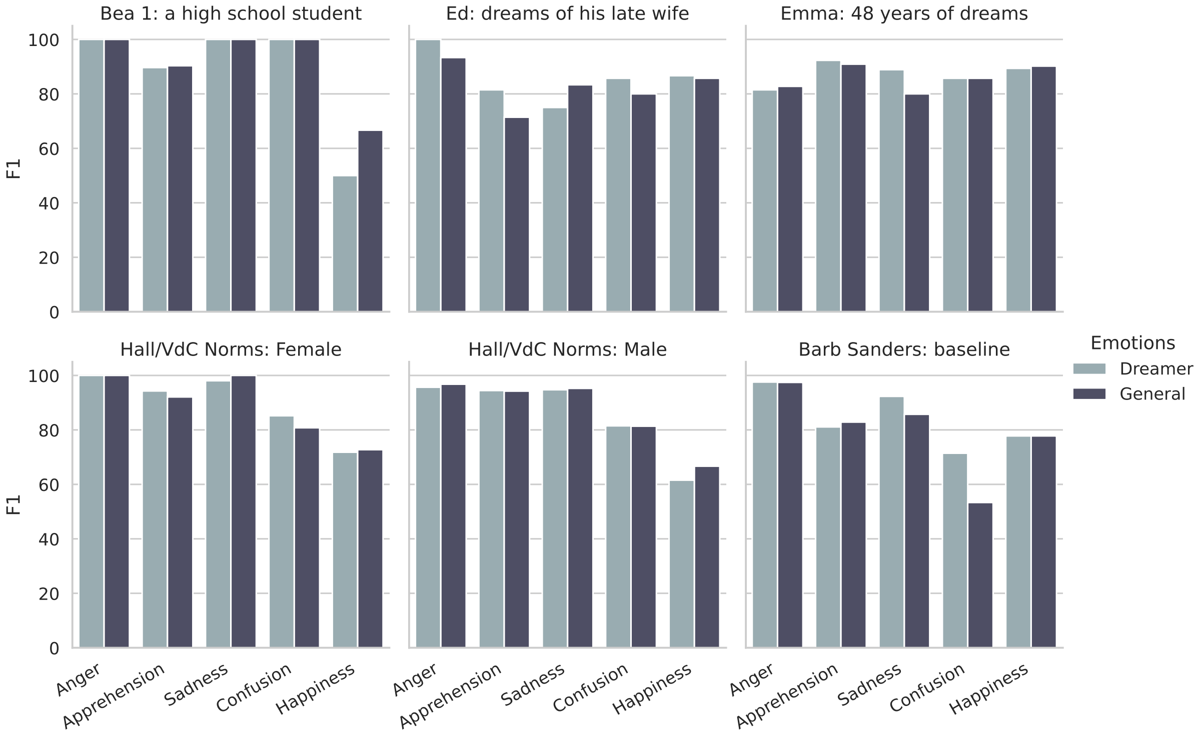

Figure 14: Single-emotion: Series and emotions analysis. Results (in terms of reference-class F1 scores) obtained by the model for each HVDC emotion (x-axis), DreamBank's series (diagrams), and Emotions (Dreamer vs. General, hue). For *happiness*, the F1 scores reference class is POSITVE, while all other HVDC emotions share NEGATIVE as their reference class for the reported F1 scores.

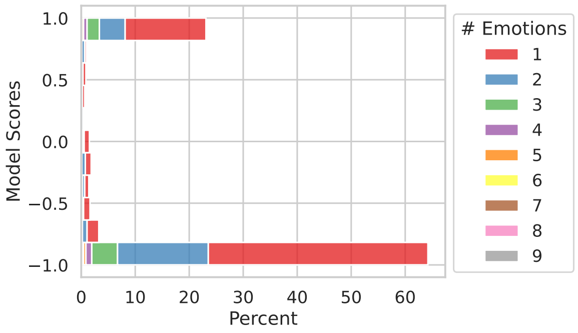

Figure 15: Model scores distribution. In-detail visualisation of the Model scores distribution, divided by the number of emotions per report, presented in Figure 11 from Section B.1.

Parameter	Value
BERT-input max-len	512
epochs	10
learning rate	0.00001
batch size	8
input truncation	True
truncation-to	max-length

Table 3: Hyper-parameters used for training the architectures in Section 4.

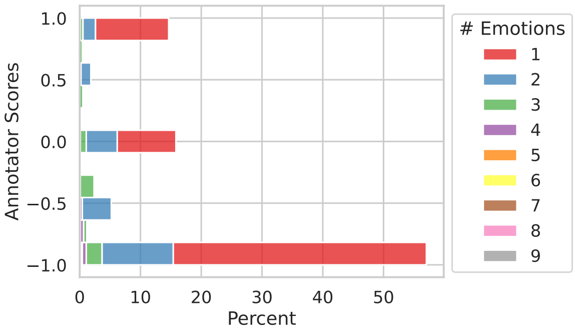

Figure 16: Annotator scores distribution. In detail visualisation of the Annotator scores distribution, divided by the number of emotions per report presented in Figure 11 from Section B.1.

C.3 Support-Score Correlation Analysis

Table 4 and Figure 18 summarise the results of the correlation analysis from Section 4.2.1. Overall, this analysis indicated no clear relationship between the number of test instances containing a specific emotion and the models' final performance in the ablation experiment.

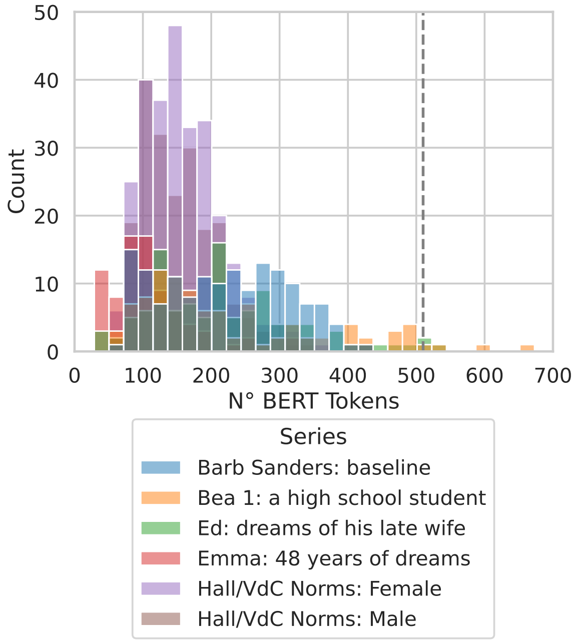

Figure 17: BERT-token distribution. Number of tokens per dream report according to the BERT tokenizer, divided by DreamBank Series. The vertical dotted line signals the indicative number of 510 tokens, after wich only 6 reports can be found.

Series	Spearman's ρ	p
Bea 1: a high school student	0.3000	0.6238
Ed: dreams of his late wife	-0.7182	0.1718
Emma: 48 years of dreams	0.7000	0.1881
Hall/VdC Norms: Female	0.7906	0.1114
Hall/VdC Norms: Male	0.7379	0.1546
Barb Sanders: baseline	0.8208	0.0886

Table 4: Correlation analysis between F1 score and support (# items) per single emotion in the ablation experiment. Each row of the table presents the results of the correlations between the number of instances containing a given emotion, and the obtained F1 scores (see Figures 3 and 18 for further visual breakdowns). Columns describe the single Series under investigation, the ρ coefficient and the p value of each correlation.

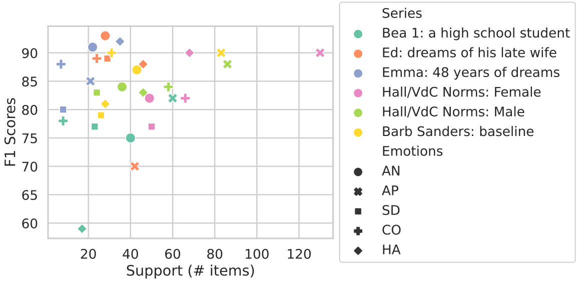

Figure 18: Ablation's experiment, score vs support correlation analysis. Visualisation of the correlation analysis, presented in Table 4, between the number of test items (x-axis) and F1 scores (y-axis), for each combination of Series and emotion.