Lexico-semantic and affective modelling of Spanish poetry: A semi-supervised learning approach

Alberto Barbado[1,2,*], María Dolores González[3,4], and Débora Carrera[1]

[1]Universidad Politécnica de Madrid, Departamento de Inteligencia Artificial, Madrid, Spain
[2]Telefónica, Madrid, Spain
[3]IFEMA, Madrid, Spain.
[4]Universidad Oberta De Catalunya, Barcelona, Spain
[*]Corresponding author email: alberto.barbadogonzalez@telefonica.com

Abstract

Text classification tasks have improved substantially during the last years by the usage of transformers. However, the majority of researches focus on prose texts, with poetry receiving less attention, specially for Spanish language. In this paper, we propose a semi-supervised learning approach for inferring 21 psychological categories evoked by a corpus of 4572 sonnets, along with 10 affective and lexico-semantic multiclass ones. The subset of poems used for training an evaluation includes 270 sonnets. With our approach, we achieve an AUC beyond 0.7 for 76% of the psychological categories, and an AUC over 0.65 for 60% on the multiclass ones. The sonnets are modelled using transformers, through sentence embeddings, along with lexico-semantic and affective features, obtained by using external lexicons. Consequently, we see that this approach provides an AUC increase of up to 0.12, as opposed to using transformers alone.

1 Introduction

Text mining techniques are useful for discovering patterns and extracting insights from within different sources of text data sets. This is useful for several tasks, such as classifying texts into opinions, sentiments, emotional states or topics, by modelling the input texts through its lexico-semantic, affective or semantic meaning (Liu 2012).

Most of the research studies focus on prose texts (Berry and Castellanos 2004), with less analyses for poetic documents, mainly because of the complexity of its meaning (like metaphors), and structure (like rhyme or enjambment) (Jakobson 1960). Another problem is that there is a lack

of poetry corpora (Kaplan and Blei 2007). Nonetheless, poetry computational analysis is growing within the literature, and studies have shown that several text mining task can be tacked through similar approaches used for prose texts (Greene, Bodrumlu, and Knight 2010; Kao and Jurafsky 2012).

Following this, poetry classification can be achieved through Machine Learning (ML), either from a supervised approach (Lou, Inkpen, and Tanasescu 2015; Jamal, Mohd, and Noah 2012; Ahmed et al. 2019; Ahmad et al. 2020), or from an unsupervised one when there is no prior information about the categories (Navarro-Colorado 2018; Ullrich et al. 2017; Barbado et al. 2021).

On many situations, the labelled data available is scarce, not being enough to train a supervised ML model, but being enough to justify using it versus following a fully unsupervised approach. Here, the literature proposes the usage of semi-supervised learning techniques (Van Engelen and Hoos 2020), which has been already proven useful for non-poetic text classification (Tian et al. 2018; Giasemidis et al. 2018; Yu et al. 2019). However, there is a lack of analyses that study the usage of semi-supervised learning approaches for poetry text classification, particularly with Spanish poems.

For any of these tasks, poems need to be modelled first through several features. This was traditionally approached by using lexical or affective information (Barros, Rodriguez, and Ortigosa 2013; Aryani et al. 2016), by word models (Jamal, Mohd, and Noah 2012; Kumar and Minz 2014), and more recently through the usage of transformers (Ahmad et al. 2020; Rosa, Ros, and González-Blanco 2020). However, poems are rich texts, and only using one type of features may not be enough to represent their meaning. However, the literature on poetic text modelling also lacks studies regarding this.

Our research focuses on these two aspects: using a labelled corpus of Spanish sonnets and follow a semi-supervised learning approach for inferring the categories of a whole poem corpus, and analyse the importance of combining both lexico-semantic and affective features with State of the Art transformers through sentence embeddings. We use the labelled corpus from (Barbado et al. 2021) in order to predict the psychological, affective and lexico-semantic categories of a corpus of 4572 sonnets.

The contributions of the paper are:

- Propose a semi-supervised learning framework that infers the psychological, lexico-semantic and affective categories of Spanish sonnets by using affective and lexico-semantic features along with sentence embeddings.

- Provide a benchmark for Spanish poetry modelling for both binary and multiclass classification. We use the 21 binary psychological categories and the 10 multiclass affective and lexico-semantic categories from "Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels" (DISCO PAL) corpus (Barbado et al. 2021).

- Analyse quantitatively the importance of combining lexico-semantic and affective features along with transformers.

- Evaluate extending the corpus from the original sonnets used for DISCO PAL (sonnets from XVth to XIXth century), to other modern sonnets (sonnets from XXth century).

2

The remainder of tha paper is organized as follows: Section 2 presents the related literature, focusing on poetry ML classification, as well as the importance of predicting these categories in poems for several domains. Section 3 describes the corpus used, as well as the methodology followed. Section 4 presents the empirical results. Finally, Section 5 shows the conclusions and highlights several research lines that can be pursued.

2 Related Work

In this Section, we present related work regarding the modelling of poems for inferring their general meaning, along with the context of our research, particularly to highlight the importance of identifying lexico-semantic, affective and psychological categories within poetry.

2.1 Poetry Classification

Poetry express concepts, themes and motifs that can be modelled through its content, similarly to what is done with other literary compositions, through text mining techniques (Liu 2012; Miall and Kuiken 1994). A common approach in the literature is through supervised techniques when there is prior knowledge about the poem themes or general meaning. This can be achieved by using Machine Learning (ML) supervised methods.

This approach appears for English poetry in (Lou, Inkpen, and Tanasescu 2015), where themes "Love", "Nature", "Social Commentaries", "Religion", "Living", "Relationships", "Activities", "Art & Sciences", and "Mythology & Folklore", are predicted through supervised ML in a corpus of 7214 poems. The authors use a Support Vector Machine (SVM) algorithm to model the relationship between Term-Frequency Inverse-Document-Frequency (TD-IDF) features, as well as Latent Dirichlet Allocation (LDA), to represent the poem content and train the ML model. This approach is similar in other languages. In (Jamal, Mohd, and Noah 2012), the authors also use a SVM to classify 1500 Malaysian poems into 10 categories. The features used for modelling the poems are also based on TF-IDF. In contrats to this approach, in (Kaur and Saini 2017) classify 240 Punjabi poems into 4 categories, benchmarking different ML supervised models (such as AdaBoost), where the input features are lexical variables. In (Ahmed et al. 2019) authors also benchamrk several ML models (SVM, Naive Bayes, and Linear Support Vector Cassification) for classifying Arabic poetry (92 poems, with a number of verses between 600 and 500) into four categories ("Love", Religious/Islamic", "Politic", "Social"). s input features, they use Boolean vector models, instead of other word counting ones. Besides inferring the category for the whole poem, other approaches focus only on a part of it. In (Kumar and Minz 2014) shows an approach for verse classification in Basque poetry with 212 strophes, considering 6 categories. As input features, they use a Bag of words (BOW) model. In (Rosa, Ros, and González-Blanco 2020), authors predict the metric pattern category of Spanish poem verses using transformers, together with a supervised learning approach. Focusing on Spanish poetry, (Barros, Rodriguez, and Ortigosa 2013) uses a supervised ML model (Decision Tree) for classifying Quevedo's poems (with a corpus of 185 poems) into 4 classes, "Love", "Lisi", "Satiric", and "Philosophical, Moral and Religious (PMR)". As input features for modelling the poems, the

authors identified words related to particular emotions (such as "joy", "sadness"...), counting their occurrences within the poems for building the features. More recent researches focus on the usage of Deep Learning (DL). In (Ahmad et al. 2020) authors use a Long Short-Term Memory (LSTM) based architecture, together with word embeddings, for predicting 13 emotional states in a corpus of 9142 poems.

The aforementioned studies either have a large labelled corpus, or they work with a small corpus of poems. However, to the best of our knowledge, there are no prior studies of semi-supervised approaches for poetry modelling, specially when we consider Spanish poems. Also, they either focus on transformers for feature generation, or they use lexico-semantic features alone. They do not consider together affective and lexico-semantic features along with transformers.

2.2 Poetry for Pedagogical and Therapeutic Purposes

Poetry therapy is the intentional application of written and spoken words for growth and healing, using a language that is condensed, replete with sensory images, and charged with meaning (Gorelick 2005). Its process involves an interaction among three key elements: the poem, the therapist, and the patient/user. The use of literature as a key element in the therapeutic context is a distinguishing characteristic of poetry therapy. (Hynes and Hynes-Berry 2019) state that the literature plays an important role in the therapeutic process by serving as a catalyst. (Hedberg 1997) highlights the cognitive benefits of poetry therapy, demonstrating the ability of poetry to communicate learning, humor, insight and improve verbalizing feeling. Literature addresses how the practice of poetry as a therapy can promote psychological well-being or happiness, based on the pillars of the so-called PERMA model (positive emotion, engagement, relationships, meaning, and accomplishment) (Seligman 2011).

Because of that, poetry therapy has been proven beneficial in several fields, such as psychology, psychiatry, or psychotherapy, as an interdisciplinary treatment method with practical applications, as studied by (LEEDY 1970) and (Silverman 1983). (Silverman 1988) analyses its use in psychotherapy and counselling, (Mcardle and Byrt 2001) explain how literature benefits mental health using expressive writing, bibliotherapy, and poetry therapy. (Lerner 1997) explore the concept "Poetry fixes me", where poetry provides healing effects. Research also exists in specific areas, as the application of poetry therapy in elderly care and terminally ill patients (Tegnér et al. 2009), where poetry can help enabling the expression of individuals' deepest unspoken concerns within the field of palliative care (Robinson 2004), as well as in other sociocultural contexts (Sharma 2020).

Poetry is also beneficial for education, being useful for enhancing learning in of all the development phases. For childhood, it helps developing language skills, while deepening the message learnt, thanks to poetic elements such as the rhyme that create an awareness of the child's own capacity in such a way that it helps to link the words with the relationship of their meaning (Samuelsson et al. 2009). For old age, poetic language allows to awaken memories related to the affections and feelings experienced by the person, providing the cognitive and affective stimuli of the brain (Aadlandsvik 2008).

With that, we see that the work carried out in this paper has relevance in several fields; having an

extensive corpus of poetry, with identified topics related to psychology, may be useful for applications like poetry therapy.

3 Method

In this Section, we present the sonnet corpus used, along with the semi-supervised modelling approach that we propose for inferring the values of several lexico-semantic, affective and psychological concepts evoked by the individual sonnets.

3.1 Corpus Description

The corpus used within this paper is a combination from three different input corpora. First, we use the corpus from DISCO (Diachronic Spanish Sonnet Corpus) (Ruiz et al. 2018), which includes 4085 sonnets from 15th to 19th centuries. This is combined with the corpus DISCO PAL (Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels) from (Barbado et al. 2021), a subset of 274 sonnets from DISCO with expert annotations for the following affective, lexico-semantic and psychological categories:

Affective categories	
Valence	Arousal
Happiness	Disgust
Anger	Sadness
Fear	

Table 1: Affective categories used with a scale of 1 to 4, using integer values.

Lexico-semantic categories	
Concreteness	Imageability
Context availability	

Table 2: Lexico-semantic categories used with a scale of 1 to 4, using integer values.

Psychological categories	
Solitude (Soledad)	Anxiety (Ansiedad)
Illusion (Ilusión)	Anger/Wrath (Ira)
Daydream (Ensoñación)	Instability (Inestabilidad)
Grandeur (Grandiosidad)	Idealization (Idealización)
Pride (Orgullo)	Depression (Depresión)
Irritability (Irritabilidad)	Disappointment (Desilusión)
Dramatisation (Dramatización)	Prejudice (Prejuicio)
Aversion/Loathing (Aversión)	Insecurity (Inseguridad)
Helplessness (Impotencia)	Vulnerability (Vulnerabilidad)
Fear (Temor)	Obsession (Obsesión)
Compulsion (Compulsión)	

Table 3: Psychological categories used with a scale of 0 to 1 (binary values).

Finally, we expand the corpus by including 532 sonnets from (Londoño 2013) with authors from the XXth century with Spanish as mother tongue. The corpora DISCO and DISCO PAL contains some multiple-part sonnets. For this paper, we will only focus on single-part sonnets (2 quartets and

2 tercets, for a total of 14 lines in 4 stanzas). This means using 270 sonnets from DISCO PAL and 4040 from DISCO. Thus, the total number of sonnets used are 4572, with 270 annotated with expert domain knowledge. Figure 1 shows the histogram of words for our new corpus, with and without stopwords (M=91.8, Std=8.7; M=43.5, Std=5.2), compared to the subset of XXth century (M=94, Std=10.3; M=41.6, Std=6.7) and to the subset annotated from DISCO PAL (M=92.2, Std=8.73; M=44.7, Std=5.0).
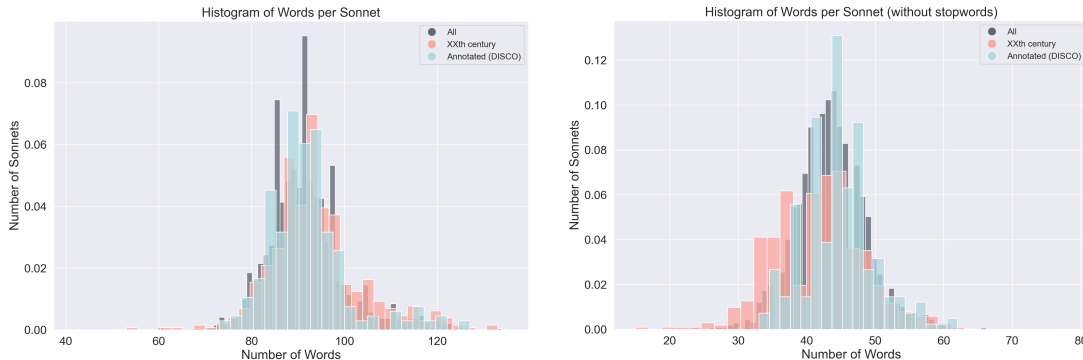

Figure 1: Histogram of words, with and without stopword, for each sonnet considering the whole corpus, the subset of new sonnets from S.XX, and the annotated subset from (Barbado et al. 2021).

Both the number of sonnets per psychological category (with "Prejudice" and "Obsession" the categories with fewer sonnets, 30 and 32 respectively), as well as the number of sonnets from XXth included, are enough from the a statistical point of view based on a power analysis with an alpha of 0.1, a Cohen's d of 0.8 and the default statistical power of 0.8 (which sets the minimum in 20) (Cohen 1992; Sullivan and Feinn 2012). The number of sonnets is also higher when compared to other previous works related to poetry, such as (Ullrich et al. 2017; Aryani et al. 2016), where the authors work with the categories "friendliness", "sadness" and "spitefulness" and they are associated to 19, 21 and 17 poems respectively. Figure 2 shows the number of sonnets per period, including the new sonnets from XXth century.

3.2   Semi-supervised Modelling

With the corpus described in Subsection 3.1, we follow the flowchart described in 3 in order to infer the categories from Tables 1 and 2 for the whole corpus using the information available at DISCO PAL.

The first part of the flowchart is a NLP (natural language processing) pipeline that performs the following steps: lemmatization and part-of-speech tagging, tokenization, elimination of stopwords and stemming.

The second phase of the flowchart includes all the steps for the lexico-semantic and affective modelling of the sonnets from the corpus, in order to infer its General Affective Meaning (GAM) (). These features are the ones from (Barbado et al. 2021) since they have been already proved
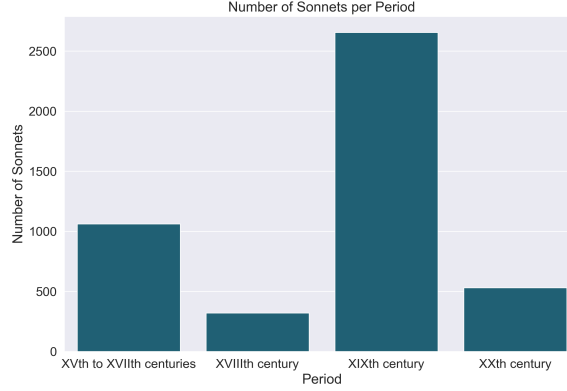
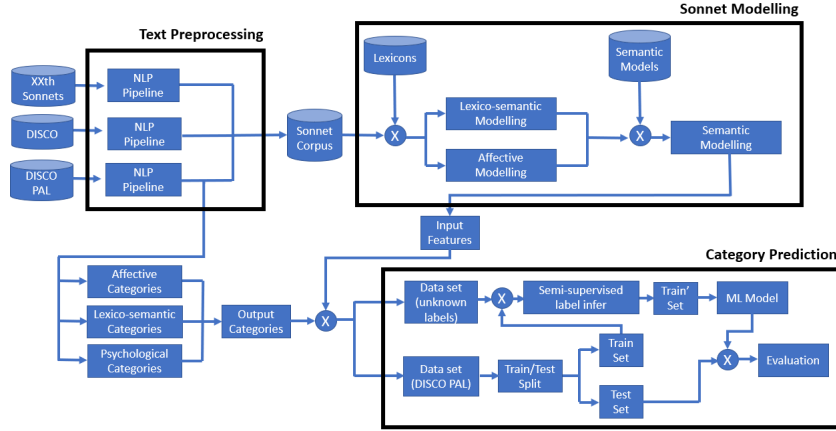Figure 2: Sonnets per time period, including the new ones for S.XX.

Figure 3: Flowchart followed by the process, with three main phases: an initial text processing, the lexico-semantic and affective modelling of the sonnets, and the category prediction using semi-supervised learning.

as useful for modelling the GAM of Spanish sonnets. They are similar to the ones used in other works, such as (Ullrich et al. 2017) for inferring the GAM on German poetry. The features are described in Annex (A.2). These features are obtained using several Spanish lexicons that provide lexico-semantic and affective values for individual words. In particular, we use (Pérez-Sánchez et al. 2021), (Stadthagen-González et al. 2018), (Hinojosa, José A et al. 2016), (Guasch, Ferré, and Fraga 2016) and (Ferré et al. 2017). These lexicons are combined into one, where we keep the mean value for each word in case there are duplicates. The lexicons do not need rescaling since a features have the same ranges (valence, arousal, concreteness, imageability and context availability features have a range of 1 to 9; happiness, anger, sadness, fear and disgust features have a range of 1 to 5).

In order to assign lexico-semantic and affective values to the individual words for building the features from (A.2), we use the stemmed words in the sonents (after the text preprocessing phase)

7

and we match them with the corresponding stemmed lexicon entry. If there are several possible entries, we assign the mean value of all of them. We use the stemmed version of the words for crossing the two data source since it provides the best coverage of sonnet words, as Table 4 shows. The total word coverage is 47%, with some words missing within the lexicon, as can be seen in Figure 4, representing the word cloud of missing the sonnets of XXth century, as well as on the full corpus. Most of the missing words are because some sonnets are written in old Spanish. This is the reason why the percentage of coverage increases when we consider only XXth. The percentage for stemmed words within this subset is 74%, and it increases to 93% if we consider the relative appearances of the words, highlighting that in many cases the missing words are hapaxes.

Lexicon	Processing	Corpus (All)	Corpus (DISCO)	Corpus (XXth)	Corpus (Annotated)
All Lexicons	Stemmed	0.47	0.48	0.74	0.73
All Lexicons	Lemmatized	0.39	0.4	0.63	0.64
All Lexicons	Original	0.2	0.21	0.34	0.36
(Pérez-Sánchez et al. 2021)	Stemmed	0.05	0.05	0.10	0.11
(Pérez-Sánchez et al. 2021)	Lemmatized	0.04	0.04	0.08	0.09
(Pérez-Sánchez et al. 2021)	Original	0.02	0.02	0.04	0.05
(Stadthagen-González et al. 2018)	Stemmed	0.37	0.37	0.57	0.55
(Stadthagen-González et al. 2018)	Lemmatized	0.26	0.26	0.38	0.37
(Stadthagen-González et al. 2018)	Original	0.14	0.14	0.21	0.21
(Hinojosa, José A et al. 2016)	Stemmed	0.05	0.05	0.11	0.12
(Hinojosa, José A et al. 2016)	Lemmatized	0.03	0.04	0.08	0.08
(Hinojosa, José A et al. 2016)	Original	0.02	0.02	0.03	0.04
(Guasch, Ferré, and Fraga 2016)	Stemmed	0.08	0.08	0.16	0.15
(Guasch, Ferré, and Fraga 2016)	Lemmatized	0.05	0.06	0.10	0.1
(Guasch, Ferré, and Fraga 2016)	Original	0.03	0.03	0.06	0.07
(Ferré et al. 2017)	Stemmed	0.12	0.13	0.23	0.25
(Ferré et al. 2017)	Lemmatized	0.09	0.09	0.16	0.18
(Ferré et al. 2017)	Original	0.05	0.05	0.10	0.11

Table 4: Percentage of words from the corpus (without stopwords) within each of the lexicons, considering original words, words after lemmatization and words after stemming.

Figure 4: Missing from all lexicons, considering either the full corpus or the subset of sonnets from XXth century.

After the lexico-semantic and affective modelling steps, the sonnet modelling phase includes a semantic modelling step based on word embedding techniques. Particularly, since sonnets are highly-

structured and short text compositions, our proposed approach is modelling them through pre-trained sentence transformers. This can be done since pre-trained models for sentence embeddings are normally fine-tuned for small text compositions (up to 128 words (Reimers and Gurevych 2020)), and within our corpus, sonnets do not exceed 65 word pieces (without stopwords). The pre-trained models that we evaluate for modelling the corpus of sonnets are the multi-lingual models shown below (Reimers and Gurevych 2019):

- *quora-distilbert-multilingual*: Model using DistilBertModel as transformer, mean pooling, with 768 features.

- *stsb-xlm-r-multilingual*: Model using XLMRobertaModel as transformer, mean pooling, with 768 features.

- *paraphrase-multilingual-mpnet-base-v2*: Model using XLMRobertaModel as transformer, mean pooling, with 768 features.

- *paraphrase-xlm-r-multilingual-v1*: Model using Distil-Roberta-Base as transformer, mean pooling, with 768 features.

- *distiluse-base-multilingual-cased-v1*: Model using the Universal Sentence Encoder, mean pooling, with 512 features.

These sentence embedding models are complemented with our proposed sentence embedding that weights the words according to their affective meaning. Particularly, we normalize the lexicon and associate as a weight for each word the maximum normalized value among all the lexico-semantic and affective features. Then, we apply a mean weighted pooling according to those weights. This is shown in Equation 1, with N the number of words, $\vec{v}(k)$ the word embedding for word k, $w(k)$ the word weight. This $w(k)$ value corresponds to the maximum normalized feature value from (A.2).

$$\vec{v}_n = \frac{\sum_{k=1}^{N} \vec{v}(k) \times w(k)}{N}$$
$$\text{with } w(k) = max(f_1(k), f_2(k), ..., k_m(k))$$
(1)

For the individual word embeddings, we use:

- *bert-base-multilingual-cased*: Pretrained multilingual based model, trained over Wikipedia texts, with 768 features (Devlin et al. 2018).

- *dccuchile/bert-base-spanish-wwm-cased*: BETO, Spanish version of BERT, which has been seen useful in some domains when compared to multilingual models (Cañete et al. 2020). 768 features.

The final phase consists in inferring the lexico-semantic, affective and psychological categories through the modelling of sonnets through lexico-semantic and affective features. (Barbado et al. 2021) already proposes using the mean values from the input features as GAM for the sonnet. However, we will see if the inferring process through semi-supervised learning improves it. In our

approach, we select a random small subset of sonnets from DISCO PAL (in particular, we sample a maximum of two sonnets per category for building the train set), using them to for the semi-supervised learning task, together with the rest of the corpus. For tackling the semi-supervised learning task, we consider the following approaches:

- Combining the algorithm from (Yarowsky 1995) together with LightGBM (Light Gradient Boosting Machine) algorithm (Ke et al. 2017). The reason behind it is that this semi-supervised learning algorithm is a self-training model that can be combined with any supervised Machine Learning algorithm, and LightGBM is an algorithm well-known for its good performance over imbalanced data sets, such as our sonnet corpus (where the annotated categories are imbalanced).

- Using the semi-supervised LabelSpreading algorithm (Zhou et al. 2004), which infers the missing labels by finding nearby points in the space and assigning them the same label as the data points with known labels. This can be done by either using a K-Nearest Neighbors (KNN) approach, or a Radial Basis Function (RBF) one.

- Using the semi-supervised learning output from LabelSpreading as a pre-training phase for training a supervised algorithm (Van Engelen and Hoos 2020). In our case, the supervised ML algorithm is LightGBM. This offers also the possibility of using over-sampling techniques over the whole training data. In our case, we will analyse the results by using SMOTE (Synthetic Minority Over-Sampling Technique) (Chawla et al. 2002)

With that, an individual semi-supervised model will be trained for predicting each of the categories mentioned in Subsection 3.1.

4   Results

In this Section, we present the results of the different experiments that we have carried out in order to answer the following research questions:

- **Q1**: *Is the proposal able to accurately infer the lexico-semantic, affective and psychological categories over the whole corpus?*

- **Q2**: *Are the lexico-semantic and affective features significantly important for predicting the different categories?*

- **Q3**: *Are the results significantly different if we use only the original DISCO sonnets (since the annotated sonnets are a subcorpus from DISCO) compared to the results obtained by using additional sonnets (XXth sonnets)?*

For **Q1**, we focus on three well-known classification metrics in order to assess the quality of the predictions.

First, we consider F1 metric. This metric is common within classification tasks, and useful when the labels are imbalanced (Jeni, Cohn, and De La Torre 2013). (Lou, Inkpen, and Tanasescu 2015)

uses binary classifiers to identify on of the 9 possible category associated to the poems. Using this as a reference benchmark, the results on F1 range from **0.744** to **0.948** on their best run (with an average of **0.835**). As a reference for Spanish language models (but for prose texts), (Alhuzali and Ananiadou 2021) shows SOTA benchmark metrics for multiclass emotion text classification. For Spanish language models, F1 is between **0.532** (F1-macro) and **0.641** (F1-micro).

We also use the Cohen's Kappa metric, that measures the agreement between one annotator (the ground truth and another annotator (the predictions) (Cohen 1960). This metric is useful for both comparing models (and choosing the one with better kappa), as well as for seeing if the results are good by themselves, since it offers thresholds to evaluate the results. (Landis and Koch 1977) propose the following ones: $kappa < 0$: Poor agreement, $0 < kappa < 0.20$: Slight agreement, $0.21 < kappa < 0.40$: Medium (moderate) agreement, $0.41 < kappa < 0.60$: Sufficient (fair) agreement, $0.61 < kappa < 0.80$: Considerable agreement, and $0.81 < kappa < 1.0$: Almost perfect agreement.

Third, we use the Area Under the Curve (AUC) as a metric to assess the quality of the explanations. Though this metric depends on the context and not every domain considers the same thresholds to identify when the values are good, on general terms the following reference values can be considered (Hosmer Jr, Lemeshow, and Sturdivant 2013): $AUC = 0.5$: No discrimination, $0.5 < AUC < 0.7$: Poor discrimination, $0.7 \geq AUC < 0.8$: Acceptable discrimination, $0.8 \geq AUC < 0.9$: Excellent discrimination and $0.9 \geq AUC$: Outstanding discrimination. Using again the reference of (Lou, Inkpen, and Tanasescu 2015) for poetry in English, the AUC ranges on their best run from **0.628** to **0.766** (with an average of **0.700**).

For **Q1**, **Q2** and **Q3** we apply a Random Cross Validation approach (CV) in order to check the results over different runs (Q1), when using affective features versus not using them (Q2), and when using only DISCO versus using also external sonnets (Q3). We use 20 random samples, since the value is enough from the a statistical point of view based on a power analysis with an alpha of 0.1, a Cohen's d of 0.8 and the default statistical power of 0.8 (which sets the minimum in 20).

4.1 Nomenclature

Here we indicate the nomenclature we use for the different combination of algorithms that we use within our analyses.

- LS-LightGBM: Combination of LabelSpreading with LightGBM. It can either be LS-LightGBM-KNN (KNN kernel) or LS-LightGBM-RBF (RBF kernel). If it includes SMOTE, then we refer to it as LS-LightGBM-SMOTE-KNN or LS-LightGBM-SMOTE-RBF.

- ST-LightGBM: Self-Training algorithm with LightGBM.

4.2 Data Description

For the analyses, we consider a Random Cross Validation (CV), where we use a subsample of sonnets for training an individual ML model on for predicting each category, and then we evaluate the results over the remaining sonnets. We choose the random subset of sonnets by selecting two sonnets from

each category value This leaves an average of 28% sonnets for training, and the remaining 72% for evaluation.

4.3 Psychological Categories

Here, we analyse Q1, Q2 and Q3 regarding the binary psychological categories. We first check Q1 by choosing the semantic transformers and semi-supervised models (introduced in Subsection 3.2) that yield the best metric results over the Random CV. Table 5 shows the best semantic and predictive models for each category, along with their mean metric values. It also shows the mean data distribution of the test set for each category, where we see that in all cases, except for 'Prejudice', the number of sonnets with that category are above the reference of 20. Regarding F1, we see that 9 categories have values above the reference 0.744 considered. For Cohen's Kappa, 9 categories have values belonging to the "moderate agreement" group; the rest of the categories have a "slight agreement". Finally, for AUC, 18 categories have values above the minimum reference of 0.628. Among those categories, 8 of them have values within the "acceptable discrimination" general reference. In particular, the following categories have at least two metrics with values above the SOTA references: "Aversion", "Depression", "Disappointment", "Dramatisation", "Daydream", "Grandeur", "Idealization", "Illusion", "Instability", "Insecurity", "Anger (binary)", "Irritability", "Pride", "Prejudice", "Solitude" and "Vulnerability" (76% of the categories).

Category	Semantic Model	Predictive Model	N Class 0 Test	N Class 1 Test	F1 (weighted)	Kappa	AUC
Anxiety	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	146.0	51.25	0.69	0.17	0.66
Aversion	stsb-xlm-r-multilingual	ST-LightGBM	130.4	66.9	0.69	0.32	0.74
Compulsion	bert-base-multilingual-cased	ST-LightGBM	159.3	38.0	0.72	0.06	0.55
Depression	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	172.5	24.8	0.83	0.18	0.71
Disappointment	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	167.3	30.0	0.78	0.13	0.64
Dramatisation	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	122.8	74.5	0.63	0.22	0.66
Daydream	quora-distilbert-multilingual	LS-LightGBM-SMOTE-KNN	164.6	32.7	0.78	0.15	0.7
Grandeur	stsb-xlm-r-multilingual	ST-LightGBM	118.5	78.8	0.67	0.31	0.74
Idealization	bert-base-spanish-wwm-cased	LS-LightGBM-SMOTE-KNN	119.3	78.0	0.66	0.3	0.7
Illusion	paraphrase-multilingual-mpnet-base-v2	LS-LightGBM-SMOTE-KNN	144.3	52.9	0.71	0.26	0.72
Helplessness	quora-distilbert-multilingual	LS-LightGBM-SMOTE-RBF	156.6	40.7	0.69	0.19	0.67
Instability	bert-base-multilingual-cased	LS-LightGBM-SMOTE-KNN	154.6	42.7	0.72	0.21	0.68
Insecurity	distiluse-base-multilingual-cased-v1	ST-LightGBM	167.7	29.6	0.79	0.08	0.65
Anger (binary)	stsb-xlm-r-multilingual	ST-LightGBM	160.8	36.5	0.75	0.17	0.68
Irritability	quora-distilbert-multilingual	ST-LightGBM	175.6	21.7	0.84	0.1	0.69
Obsession	quora-distilbert-multilingual	LS-LightGBM-SMOTE-RBF	176.2	21.1	0.82	0.08	0.61
Pride	stsb-xlm-r-multilingual	ST-LightGBM	144.0	53.3	0.7	0.23	0.69
Prejudice	quora-distilbert-multilingual	ST-LightGBM	178.9	18.4	0.87	0.1	0.74
Solitude	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	154.8	42.5	0.75	0.24	0.72
Fear (binary)	bert-base-multilingual-cased	LS-LightGBM-SMOTE-KNN	131.5	65.8	0.6	0.15	0.62
Vulnerability	bert-base-multilingual-cased	LS-LightGBM-SMOTE-KNN	107.8	89.5	0.59	0.21	0.67

Table 5: Best combinations of semantic models and semi-supervised techniques for the different psychological categories.

Figure 5 shows the comparison of the AUC value with respect to a baseline LigthGBM model (with and without SMOTE) without using semi-supervised learning. Considering these baseline models, in all the categories, the semi-supervised learning approach is either similar or better than the baselines counterparts. For "Dramatisation", "Obsession", "Depression", "Helplessness", "Anxiety", "Instability", "Idealization", "Fear (binary)", "Vulnerability", and "Illusion", the results are statistically different according to Wilcoxcon signed-rank test (Conover 1998), with the semi-

supervised approach being better. For the remaining categories, even though the semi-supervised approach improves the baseline models in some cases, the results are not statistically different: "Compulsion" (p-value 0.97), "Disappointment" (p-value=0.96), "Solitude" (p-value=0.89), "Insecurity" (p-value=0.78), "Grandeur" (p-value=0.43), "Pride" (p-value=0.33), "Irritability" (p-value=0.3), "Anger (binary)" (p-value=0.29), "Prejudice" (p-value=0.16), "Daydream" (p-value=0.12), and "Aversion" (p-value=0.11). The results of the semi-supervised models are also, for all the categories, significantly better than those of a random classifier (AUC=0.5).
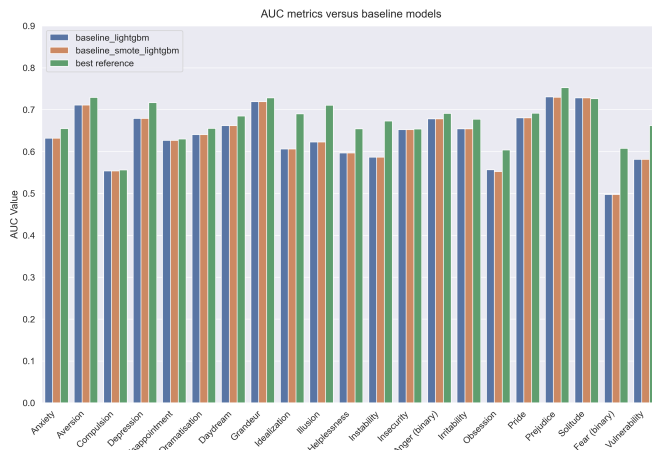

Figure 5: Comparison of the semi-supervised approach versus baseline models.

Regarding Q2 and Q3, the results are shown in Figure 6. For Q2, there are 11 categories with statistically significant differences according to Wilcoxcon signed-rank test. They are "Anxiety", "Aversion", "Daydream", "Grandeur", "Idealization", "Illusion", "Helplessness", "Instability", "Irritability", "Fear (binary)", and "Vulnerability". In all those cases, the AUC metrics improve when using the affective and lexico-semantic additional features together with the semantic transformers. The AUC improvement in those cases span from 0.02 to 0.12 (with a mean improvement of 0.07). For Q3, there are only 5 categories where the AUC metrics are statistically significant: "Aversion", "Depression", "Helplessness", "Irritability" and "Prejudice". However, using additional sonnets from the XXth is actually improving the results in all those cases. The AUC improvement in those cases span from 0.02 to 0.05 (with a mean improvement of 0.03).

4.4   Lexico-semantic and Affective Features

Similar to the previous section, here we analyse Q1, Q2 and Q3 considering the lexico-semantic and affective categories. In this case, the models are multiclass, since the categories have 4 possible values. Table 6 shows the best semantic and predictive models for each category, along with their mean metric values. It also shows the mean data distribution of the test set for each category. Regarding F1, we see that only 3 out of the 10 categories have a F1 value above the reference of 0.532 ("anger (ordinal)", "fear (ordinal)", "happiness"). For Cohen's Kappa, only "fear (ordinal)"
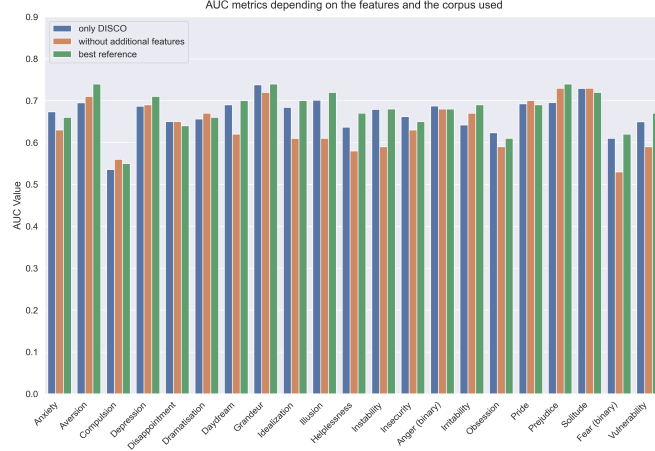
Figure 6: Comparison of the results without using affective and lexico-semantic additional features, and also when using only DISCO.

has a value within the "moderate agreement" group (the remaining categories are within the "slight agreement" one). Finally, for AUC, 6 out of the 10 categories have values above the 0.628 threshold considered (we also use as a reference the AUC used from binary classification models). In particular, the following categories have at least two metrics with values above the SOTA reference: "anger (ordinal)", "fear (ordinal)", and "happiness". "disgust", "sadness" and "valence" have metrics that only have one metric above the SOTA reference (for AUC). "arousal", "concreteness", "context availability", and "imageability" are the categories with worst results, being below the reference values for all the metrics.

Category	N Class 1 Test	N Class 2 Test	N Class 3 Test	N Class 4 Test	Semantic Model	Predictive Model	F1 (weighted)	Kappa	AUC
anger (ordinal)	144.2	32.9	18.5	5.3	stsb-xlm-r-multilingual	LS-LightGBM-SMOTE-KNN	0.63	0.14	0.65
arousal	32.0	98.6	60.3	10.0	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	0.43	0.09	0.59
concreteness	78.4	54.2	45.0	23.2	quora-distilbert-multilingual	ST-LightGBM	0.32	0.06	0.55
context availability	89.0	57.6	36.4	17.9	paraphrase-xlm-r-multilingual-v1	LS-LightGBM-RBF	0.36	0.07	0.55
disgust	108.6	55.9	32.4	4.0	stsb-xlm-r-multilingual	ST-LightGBM	0.5	0.16	0.66
fear (ordinal)	143.9	46.8	9.4	1.0	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	0.68	0.22	0.7
happiness	134.7	41.0	20.1	5.1	stsb-xlm-r-multilingual	LS-LightGBM-KNN	0.58	0.16	0.67
imageability	91.9	54.5	39.7	14.9	paraphrase-xlm-r-multilingual-v1	LS-LightGBM-RBF	0.38	0.07	0.56
sadness	67.4	56.1	55.7	21.7	paraphrase-multilingual-mpnet-base-v2	ST-LightGBM	0.4	0.2	0.67
valence	9.8	61.5	69.8	59.8	stsb-xlm-r-multilingual	ST-LightGBM	0.42	0.19	0.65

Table 6: Best combinations of semantic models and semi-supervised techniques for the different lexico-semantic and affective categories.

We also include here a comparison of the AUC value with respect to a baseline LigthGBM model (with and without SMOTE) without using semi-supervised learning in Figure 7. In all the situations where there are significant differences (according to Wilcoxcon signed-rank test), the semi-supervised approach surpasses the baseline models. The only cases where there are no significant differences are "anger (ordinal)" with SMOTE baseline (p-value=0.57), "arousal" (p-value=0.17 for LightGBM, p-value=0.89 for SMOTE variant), "concreteness" (p-value=0.11), "disgust" with SMOTE baseline (p-value=0.34), and "fear (ordinal)" (p-value=0.32). Thus, the semi-supervised

approach significantly improves the results for "context availability", "happiness", "sadness" and "valence". The results of the semi-supervised models are also, for all the categories, significantly better than those of a random classifier (AUC=0.5).
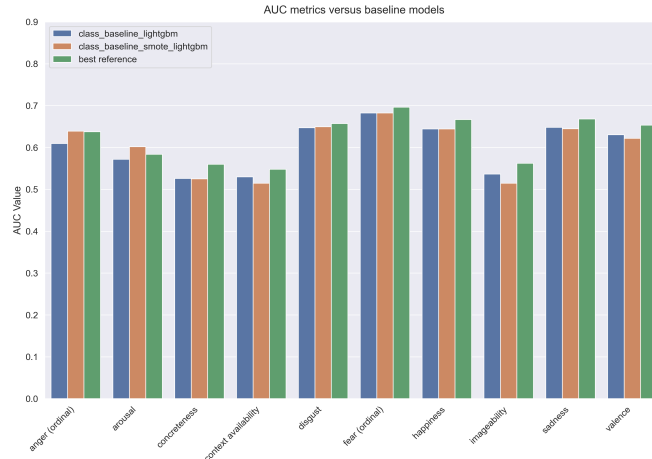

Figure 7: Comparison of the semi-supervised approach versus baseline models.

Regarding Q2 and Q3, the results are shown in Figure 8. For Q2, there are 7 categories with statistically significant differences according to Wilcoxcon signed-rank test. They are "anger (ordinal)", "arousal", "context availability", "happiness", "imageability", "sadness", and "valence"". In all those cases, the AUC metrics improve when using the affective and lexico-semantic additional features together with the semantic transformers. The AUC improvement in those cases span from 0.02 to 0.06 (with a mean improvement of 0.03). For Q3, there are only 5 categories where the AUC metrics are statistically significant: "arousal", "disgust", and "valence". Using additional sonnets from the XXth is actually improving the results for "arousal" and "valence" (improvement on AUC of 0.02). For "disgust", using XXth sonnets worsens the AUC metric lowering it on 0.01.

4.5   Limitations of our Approach

There are several limitations within our approach. The first one is that the semi-supervised learning training and evaluation uses 270 sonnets from DISCO PAL. It would be interesting to perform it over a bigger corpus of annotated sonnets. However, we think that the corpus size is big enough for carrying out these analyses and obtain statistically meaningful results, since our corpus is has more poems per independent category than other corpora from the literature (Ullrich et al. 2017), (Aryani et al. 2016), (Jacobs et al. 2017), (Obermeier et al. 2013), (Haider et al. 2020). Also, the threshold found using the statistical power serves as another reason to support the statistical analyses carried out.

Another limitation is that the analysis is applied only over highly-structured poems (sonnets) in Castilian. The results may vary if other languages or other types of poems are considered.
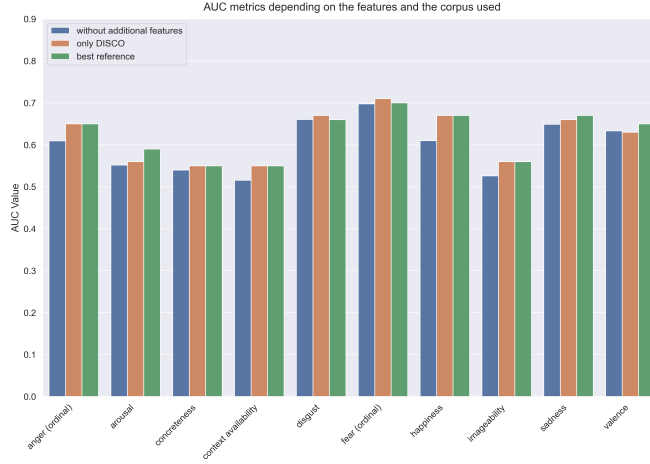
Figure 8: Comparison of the results without using affective and lexico-semantic additional features, and also when using only DISCO.

4.6 Software Used

The main libraries used for the work done in this paper are the following:

- Semi-supervised models (Buitinck et al. 2013)

- Transformers (Reimers and Gurevych 2019, Reimers and Gurevych 2020, Devlin et al. 2018, Cañete et al. 2020)

The source code and results for this paper are available in the following repository (Barbado 2021).

5 Conclusion and Future Work

In this final Section we present some potential future research lines following this work, as well as a summary of the conclusions achieved.

5.1 Future Work

In this paper we present a benchmark on supervised modeling of Spanish poetry for inferring the GAM (affective categories), together with lexico-semantic and psychological categories evoked by the poems. Our approach uses semi-supervised learning, since the labelled data set is small. However, there are other alternatives, such as one-shot learning, that could also be used for the supervised modelling task within this context.

It would also be interesting to analysis the results with other types of poems, even combining them within the same corpus with the sonnets. Also, since the semantic transformers are multilingual, and

there are available lexicons for lexico-semantic features, the approach could be potentially assessed using poems from other languages.

Finally, we use as semantic transformers models already pre-trained, but it could be interesting to evaluate the results by fine-tuning the models for poetic texts, since poetry uses linguistic elements, such as metaphors, that may not be properly represented with those models.

5.2 Conclusions

This article presents a methodology for inferring the General Affective Meaning (GAM) of Spanish sonnets, along with other lexico-semantic and psychological concepts evoked by the poems. Some of these concepts have binary values, while others are multiclass. We used a labelled subset of sonnets (270) from a bigger sonnet corpus (4572), and we applied a semi-supervised learning approach in order to model the sonnets using as few labelled sonnets as possible, in order to evaluate the results in an statistically significant subset. The sonnets are modelled using sentence transformers, together with lexico-semantic and affective features derived from their individual words by using external lexicons.

According to similar works from the literature, we saw that our proposal achieved good results for at least 76% of the psychological concepts, and for 60% of the afective and lexico-semantic ones.

We also checked that the results are significantly better by combining the features derived from the lexicons, together with the sentence transformers, showing how there is information about the sonnets that is not captured by the word embeddings.

Finally, we saw that the prediction of those concepts could be extended for additional sonnets that do not belong to the original corpus used. We included sonnets from other time periods (XXth), and checked how the results are statistically similar.

5.3 CRediT authorship contribution statement

Alberto Barbado: Conceptualization, Investigation, Writing - original draft, Writing - review and editing, Visualization, Formal Analysis, Methodology. **María Dolores González**: Writing - review and editing. **Débora Carrera**: Writing - review and editing.

References

Aadlandsvik, Ragna (2008). "The second sight: Learning about and with dementia by means of poetry". In: *Dementia* 7.3, pp. 321–339.

Ahmad, Shakeel et al. (2020). "Classification of poetry text into the emotional states using deep learning technique". In: *IEEE Access* 8, pp. 73865–73878.

Ahmed, Munef Abdullah et al. (2019). "The classification of the modern arabic poetry using machine learning". In: *Telkomnika* 17.5, pp. 2667–2674.

Alhuzali, Hassan and Sophia Ananiadou (2021). "SpanEmo: Casting Multi-label Emotion Classification as Span-prediction". In: *arXiv preprint arXiv:2101.10038*.

Aryani, Arash et al. (2016). "Measuring the basic affective tone of poems via phonological saliency and iconicity." In: *Psychology of Aesthetics, Creativity, and the Arts* 10.2, p. 191.

Barbado, Alberto (2021). *Semi-supervised learning for Spanish poetry classification.* `https://github.com/AlbertoBarbado/poetry-spanish-semisupervised`.

Barbado, Alberto et al. (2021). "DISCO PAL: Diachronic Spanish Sonnet Corpus with Psychological and Affective Labels". In: *Language Resources and Evaluation*.

Barros, Linda, Pilar Rodriguez, and Alvaro Ortigosa (2013). "Automatic Classification of Literature Pieces by Emotion Detection: A Study on Quevedo's Poetry". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, pp. 141–146.

Berry, Michael W and Malu Castellanos (2004). "Survey of text mining". In: *Computing Reviews* 45.9, p. 548.

Buitinck, Lars et al. (2013). "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122.

Cañete, José et al. (2020). "Spanish Pre-Trained BERT Model and Evaluation Data". In: *PML4DC at ICLR 2020*.

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1, pp. 37–46.

— (1992). "A power primer." In: *Psychological bulletin* 112.1, p. 155.

Conover, William Jay (1998). *Practical nonparametric statistics*. Vol. 350. John Wiley & Sons.

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: `1810.04805`. URL: `http://arxiv.org/abs/1810.04805`.

Ferré, Pilar et al. (2017). "Moved by words: Affective ratings for a set of 2,266 Spanish words in five discrete emotion categories". In: *Behavior research methods* 49.3, pp. 1082–1094.

Giasemidis, Georgios et al. (2018). "A semi-supervised approach to message stance classification". In: *IEEE Transactions on Knowledge and Data Engineering* 32.1, pp. 1–11.

Gorelick, Kenneth (2005). "Poetry therapy". In: *Expressive therapies*, pp. 117–140.

Greene, Erica, Tugba Bodrumlu, and Kevin Knight (2010). "Automatic analysis of rhythmic poetry with applications to generation and translation". In: *Proceedings of the 2010 conference on empirical methods in natural language processing*, pp. 524–533.

Guasch, Marc, Pilar Ferré, and Isabel Fraga (2016). "Spanish norms for affective and lexico-semantic variables for 1,400 words". In: *Behavior research methods* 48.4, pp. 1358–1369.

Haider, Thomas et al. (2020). "PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry". In: *arXiv preprint arXiv:2003.07723*.

Hedberg, Thomas M (1997). "The re-enchantment of poetry as therapy". In: *The Arts in psychotherapy* 24.1, pp. 91–100.

Hinojosa, José A et al. (2016). "Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions". In: *Behavior research methods* 48.1, pp. 272–284.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant (2013). *Applied logistic regression*. Vol. 398. John Wiley & Sons.

Hynes, Arleen McCarty and Mary Hynes-Berry (2019). *Bibliotherapy the interactive process: A handbook*. Routledge.

Jacobs, Arthur M et al. (2017). "What is in the brain that ink may character: A quantitative narrative analysis of Shakespeare s 154 sonnets for use in (Neuro-) cognitive poetics". In: *Scientific Study of Literature* 7.1, pp. 4–51.

Jakobson, Roman (1960). "Linguistics and poetics". In: *Style in language*. MA: MIT Press, pp. 350–377.

Jamal, Noraini, Masnizah Mohd, and Shahrul Azman Noah (2012). "Poetry classification using support vector machines". In: *Journal of Computer Science* 8.9, p. 1441.

Jeni, László A, Jeffrey F Cohn, and Fernando De La Torre (2013). "Facing imbalanced data–recommendations for the use of performance metrics". In: *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, pp. 245–251.

Kao, Justine and Dan Jurafsky (2012). "A computational analysis of style, affect, and imagery in contemporary poetry". In: *Proceedings of the NAACL-HLT 2012 workshop on computational linguistics for literature*, pp. 8–17.

Kaplan, David M and David M Blei (2007). "A computational approach to style in American poetry". In: *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. IEEE, pp. 553–558.

Kaur, Jasleen and Jatinderkumar R Saini (2017). "Punjabi poetry classification: the test of 10 machine learning algorithms". In: *Proceedings of the 9th International Conference on Machine Learning and Computing*, pp. 1–5.

Ke, Guolin et al. (2017). "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems*, pp. 3146–3154.

Kumar, Vipin and Sonajharia Minz (2014). "Poem classification using machine learning approach". In: *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*. Springer, pp. 675–682.

Landis, J Richard and Gary G Koch (1977). "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers". In: *Biometrics*, pp. 363–374.

LEEDY, JACK J (1970). "The value of poetry therapy". In: *American Journal of Psychiatry* 126.8, pp. 1183–1184.

Lerner, Arthur (1997). "A look at poetry therapy". In: *The Arts in Psychotherapy* 24.1, pp. 81–89.

Liu, Bing (2012). "Sentiment analysis and opinion mining". In: *Synthesis lectures on human language technologies* 5.1, pp. 1–167.

Londoño, Graciela Henao (2013). *A Media Voz*. URL: http://amediavoz.com/ (visited on 08/13/2021).

Lou, Andres, Diana Inkpen, and Chris Tanasescu (2015). "Multilabel subject-based classification of poetry". In: *The Twenty-Eighth International Flairs Conference*.

Mcardle, Susan and Richard Byrt (2001). "Fiction, poetry and mental health: expressive and therapeutic uses of literature". In: *Journal of Psychiatric and Mental Health Nursing* 8.6, pp. 517–524.

Miall, David S and Don Kuiken (1994). "Beyond text theory: Understanding literary response". In: *Discourse processes* 17.3, pp. 337–352.

Navarro-Colorado, Borja (2018). "On poetic topic modeling: extracting themes and motifs from a corpus of spanish poetry". In: *Frontiers in Digital Humanities* 5, p. 15.

Obermeier, Christian et al. (2013). "Aesthetic and emotional effects of meter and rhyme in poetry". In: *Frontiers in psychology* 4, p. 10.

Pérez-Sánchez, Miguel Ángel et al. (2021). "EmoPro–Emotional prototypicality for 1286 Spanish words: Relationships with affective and psycholinguistic variables". In: *Behavior Research Methods*, pp. 1–19.

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

— (Nov. 2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: https://arxiv.org/abs/2004.09813.

Robinson, Ann (2004). "A personal exploration of the power of poetry in palliative care, loss and bereavement". In: *International Journal of Palliative Nursing* 10.1, pp. 32–39.

Rosa, Javier de la, Salvador Ros, and Elena González-Blanco (2020). "Predicting metrical patterns in Spanish poetry with language models". In: *arXiv preprint arXiv:2011.09567*.

Ruiz, Pablo et al. (2018). "DISCO: Diachronic Spanish Sonnet Corpus". In: *Digital Humanities im deutschprachigen Raum*, pp. 394–397.

Samuelsson, Ingrid Pramling et al. (2009). "The art of teaching children the arts: Music, dance and poetry with children aged 2–8 years old". In: *International journal of early years education* 17.2, pp. 119–135.

Seligman, ME (2011). "Flourish: a visionary new understanding of happiness and well-being". In: *Policy* 27.3, pp. 60–1.

Sharma, Daneshwar (2020). "Writing poems: a waste of time or a savior of life–an autoethnographic exploration". In: *Journal of Poetry Therapy* 33.3, pp. 164–178.

Silverman, Hirsch Lazaar (1983). "Poetry as a psychotherapeutic intervention". In: *Psychotherapy in Private Practice* 1.2, pp. 47–64.

— (1988). "The creative patient: The use of poetry in psychotherapy". In: *The psychotherapy patient* 4.1, pp. 111–122.

Stadthagen-González, Hans et al. (2018). "Norms for 10,491 Spanish words for five discrete emotions: Happiness, disgust, anger, fear, and sadness". In: *Behavior research methods* 50.5, pp. 1943–1952.

Sullivan, Gail M and Richard Feinn (2012). "Using effect size or why the P value is not enough". In: *Journal of graduate medical education* 4.3, p. 279.

Tegnér, Ingrid et al. (2009). "Evaluating the use of poetry to improve well-being and emotional resilience in cancer patients". In: *Journal of Poetry therapy* 22.3, pp. 121–131.

Tian, Ran et al. (2018). "Semi-supervised Learning with Multi-Domain Sentiment Word Embeddings". In.

Ullrich, Susann et al. (2017). "On the relation between the general affective meaning and the basic sublexical, lexical, and inter-lexical features of poetic texts. A case study using 57 Poems of HM Enzensberger". In: *Frontiers in psychology* 7, p. 2073.

Van Engelen, Jesper E and Holger H Hoos (2020). "A survey on semi-supervised learning". In: *Machine Learning* 109.2, pp. 373–440.

Yarowsky, David (1995). "Unsupervised word sense disambiguation rivaling supervised methods". In: *33rd annual meeting of the association for computational linguistics*, pp. 189–196.

Yu, Jifang et al. (2019). "CVAE-Attention: CVAE based Semi-Supervised Sentiment Classification using Attention". In: *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence*, pp. 68–75.

Zhou, Dengyong et al. (2004). "Learning with local and global consistency". In: *Advances in neural information processing systems*, pp. 321–328.

6 Annex

6.1 A.1 - Lexico-semantic and affective features

- *valence_mean*: mean of valence mean values for the individual words.

- *valence_sd*: mean of valence standard deviation values for the individual words.

- *arousal_mean*: mean of arousal mean values for the individual words.

- *arousal_sd*: mean of arousal standard deviation values for the individual words.

- *happiness_mean*: mean of happiness mean values for the individual words.

- *happiness_sd*: mean of happiness standard deviation values for the individual words.

- *anger_mean*: mean of anger mean values for the individual words.

- *anger_sd*: mean of anger standard deviation values for the individual words.

- *sadness_mean*: mean of sadness mean values for the individual words.

- *sadness_sd*: mean of sadness standard deviation values for the individual words.

- *fear_mean*: mean of fear mean values for the individual words.

- *fear_sd*: mean of fear standard deviation values for the individual words.

- *disgust_mean*: mean of disgust mean values for the individual words.

- *disgust_sd*: mean of disgust standard deviation values for the individual words.

- *concreteness_mean*: mean of concreteness mean values for the individual words.

- *concreteness_sd*: mean of concreteness standard deviation values for the individual words.

- *imageability_mean*: mean of imageability mean values for the individual words.

- *imageability_sd*: mean of imageability standard deviation values for the individual words.

- *cont_ava_mean*: mean of context availability mean values for the individual words.

- *cont_ava_sd*: mean of context availability standard deviation values for the individual words.

- *max_arousal*: maximum value of arousal mean values for the individual words.

- *min_arousal*: minimum value of arousal mean values for the individual words.

- *max_valence*: maximum value of valence mean values for the individual words.

- *min_valence*: maximum value of valence mean values for the individual words.

- *arousal_span*: $max\_arousal - min\_arousal$

- *valence_span*: $max\_valence - min\_valence$

- *CorAro*: Spearman's correlation between the arousal mean value of the words and their position in the sonnet

- *CorVal*: Spearman's correlation between the valence mean value of the words and their position in the sonnet

- *AbsCorAro*: absolute value of CorAro

- *AbsCorVal*: absolute value of CorVal

- *sigma_aro*: $\frac{arousal\_mean}{1/\sqrt{N}}$ with N the number of words in the sonnet.

- *sigma_val*: $\frac{valence\_mean}{1/\sqrt{N}}$ with N the number of words in the sonnet.