The Sentiment of Latin Poetry. Annotation and Automatic Analysis of the Odes of Horace

Rachele Sprugnoli [*]
Università di Parma

Francesco Mambrini [**]
Università Cattolica del Sacro Cuore

Marco Passarotti [**]
Università Cattolica del Sacro Cuore

Giovanni Moretti [**]
Università Cattolica del Sacro Cuore

During the recent years, an always growing number of linguistic resources and automatic systems for sentiment analysis have been developed covering a wide range of languages. However, research in this field is still not much explored for texts written in Classical languages. Working on such languages means dealing with peculiar textual genres such as philosophical, historical or religious treatises, epic narratives, plays and poems. Poems are particularly suitable for sentiment analysis because they tell us about emotions and passions. In this paper, we describe the creation of the first small gold standard of Latin made of poems written by Horace and manually annotated with emotion polarity, but we also report about the results of a set of automatic classification experiments. In particular, we test both a lexicon-based approach, which uses a Latin polarity lexicon called LatinAffectus, and a zero-shot transfer method. We provide details about the methodology adopted for the annotation of the gold standard, the creation of LatinAffectus, the development of our experiments and we give details about the results and the limitations of the proposed approaches.[0][1]

1. Introduction

"Sentiment analysis" (SA) is often used an umbrella term to identify a set of tasks dealing with the analysis of people's opinions, attitudes, emotions, evaluations towards events or entities of different kinds (Liu 2020). These tasks are usually performed for purposes such as monitoring contents of social network or evaluating customer experience, by analysing texts like tweets and reviews. A still under-investigated yet promising research area where developing and applying SA resources and techniques is the study of literary texts written in historical and, particularly, Classical languages (e.g.

[*] Dipartimento di Discipline Umanistiche, Sociali e delle Imprese Culturali - Via D'Azeglio 85, 43125 Parma. E-mail: rachele.sprugnoli@unipr.it

[**] CIRCSE Research Centre - Largo Agostino Gemelli 1, 20123 Milano, Italy. E-mail: {francesco.mambrini,marco.passarotti,giovanni.moretti}@unicatt.it

[1] This paper is the result of the collaboration between the four authors. For the specific concerns of the Italian academic attribution system: Rachele Sprugnoli is responsible for Sections 2, 3, 4.1, 5; Francesco Mambrini is responsible for Section 4.2; Marco Passarotti is responsible for Sections 1 and 7; Giovanni Moretti developed the zero-shot classification script. Section 6 was collaboratively written by Francesco Mambrini and Rachele Sprugnoli.

Ancient Greek and Latin). While investigating the lexical properties of Classical literary texts is a century-long common practice, such investigation can nowadays (i) lead to replicable results, (ii) benefit from techniques developed for analysing the sentiment conveyed by any type of text and (iii) be performed with freely available lexical and textual resources.

Replicability of results represents a methodological turn in the Humanities, and particularly in a sector like literary criticism, which is highly based on the expert intuition of scholars that yet applies on the empirical evidence provided by textual data. One further methodological innovation in this sector comes from the interdisciplinary reuse of techniques originally developed in the area of Natural Language Processing (NLP) for the automatic analysis of genres of texts very different from literary ones, leading to a potential cross-fertilization of both fields. Finally, several state-of-the-art digital linguistic resources were built for Classical languages during the last two decades, including treebanks, WordNets and different kinds of lexicons (like for instance, valency and derivational lexicons) (Sprugnoli and Passarotti 2020). In such context, for what concerns SA, we recently built a polarity lexicon for Latin nouns and adjectives, called *LatinAffectus*. The current version of the lexicon includes around 5,000 Latin lemmas with their corresponding prior polarity value (Sprugnoli et al. 2020b). *LatinAffectus* was developed in the context of the *LiLa: Linking Latin* ERC project (2018-2023)[2] (Passarotti et al. 2020), which has built a Knowledge Base of interlinked linguistic resources for Latin based on the Linked Data paradigm, i.e. a collection of several datasets described using the same vocabulary for knowledge description and linked together. *LatinAffectus* is connected to the LiLa Knowledge Base, thus making it interoperable with the other linguistic resources linked so far to LiLa (Sprugnoli et al. 2020).

In this paper, we describe the use of *LatinAffectus* to perform the automatic polarity classification of sentences in the *Odes* (*Carmina*) by Horace (65 - 8 BCE). Written between 35 and 13 BCE, the *Odes* are a collection of lyric poems in four books. Following the models of Greek lyrical poets like Alcaeus, Sappho, and Pindar, the *Odes* cover a wide range of topics related to the individual and social life in Rome during the age of Augustus, like love, friendship, religion, morality, patriotism, the uncertainty of life, the cultivation of tranquility and the observance of moderation. In spite of a rather lukewarm initial reception, the *Odes* quickly became a capital source of influence, in particular as a model of authorial voice and identity, and a cornerstone for the definition of "lyric" poetry.[3] Considering not only the importance of the *Odes* in the history of Latin and European literature, but also the diversity of the contents and tones of the poems collected therein, we argue that analysing such work can lead to interesting results and might represent a use case to open a discussion about the pros and cons of applying sentiment and emotion analysis techniques and resources to literary texts written in ancient languages. Our experiments highlight also important methodological questions about the relation between data and interpretation in computationally-based literary criticism.

The paper is organised as follows. Section 2 describes related work, with a specific focus on the field of computational literary studies. Section 3 details the process of creation of a small Gold Standard consisting of eight randomly selected odes of Horace annotated with the emotion polarity conveyed by each of their sentences. Sections 4 and 5 present the experiments that we performed, respectively pursuing a lexicon-

[2] `https://lila-erc.eu`

[3] For an orientation on the vast subject of the fortune and reception of the *Odes*, see Baldo (2012).

based approach (using *LatinAffectus*) and a zero-shot cross-lingual classification method. Section 6 discusses the results. Finally, Section 7 concludes the paper, presenting a number of open issues and sketching the future work.[4]

2. Related Work

The growth of interest in sentiment and emotion analysis goes hand in hand with the increasing diffusion of online reviews, forums, microblogs and social networks that provide researchers with a huge volume of subjective texts, in which users express their opinions, emotions and evaluations. Sentiment and emotion analysis is also considered a valid tool in business, communication and social science studies: many applications are developed with the aim of monitoring customers' opinion towards a service or product, or to study the attitude of users on social networks (Pozzi et al. 2016). An additional line of research sits in the field of computational literary studies in which quantitative text analysis methods are used to answer questions related to literary theory and literary history (Herrmann, Jacobs, and Piper 2021). In this context, the texts that are analysed are fairy tales (Volkova et al. 2010; Mohammad 2012), novels (Zehe et al. 2016), theater plays (Schmidt and Burghardt 2018) and poems.[5] However, annotated datasets of poems and automatic systems specifically designed for poetry are not numerous. As for the datasets, Table 1 presents an overview of the available resources, which differ from each other from at least five points of view:

- type of annotators involved (experts or crowd workers): for example, the annotators of `PERC` (*Poem Emotion Recognition Corpus*) (Sreeja and Mahalakshmi 2016) were recruited through Facebook, while for building `PO-EMO` (Haider et al. 2020) both expert and crowd annotations were employed. In all the other cases, only experts, such as university students with a background in linguistics or literature, or members of poetry associations, were recruited;
- unit of annotation (from the single line to the whole poem): `DISCO PAL` (Barbado et al. 2022), `Kabithaa` (Mohanty, Mishra, and Mamidi 2018), `Kāvi` (Saini and Kaur 2020) and `PERC` are annotated at poem level, whereas labels are assigned at each line in `Poem` (Sheng and Uthus 2020) and the first book of *Iliad* (Pavlopoulos, Xenos, and Picca 2022). Multiple levels of annotation, i.e. line, stanza and poem, were instead taken into consideration in `PO-EMO` and `THU-FSPC` (*Tsinghua University-Fine-grained Sentimental Poetry Corpus*) (Chen et al. 2019). The only annotation at sentence level, that is an intermediate unit of analysis, is reported by Yeruva et al. (2020) on Aeschylus's tragedies;
- granularity of classification (from binary classes to wide sets of emotions): for example, `Kabithaa` has only two labels (positive and negative), while both `Kāvi` and `PERC` are based on the Indian concept of *Navrasa* that distinguishes nine emotions, both positive, such as *shaanti* (meaning 'peace'), and negative, such as *raudra* (meaning 'anger');

[4] The data presented in this paper are publicly released here: `https://github.com/CIRCSE/Latin_Sentiment\_Analysis`.

[5] For a survey on sentiment and emotion analysis applied to literature, see Kim and Klinger (2018).

- perspective (annotation of the emotions as intended by the author or as perceived by the reader): both approaches are covered by the available datasets. It is interesting to notice that during a preliminary annotation of the *Iliad*, both perspectives were taken into consideration and annotated by two different groups. However, the annotation of the emotions that the poet tried to provoke to the reader registered a very low inter-annotator agreement (0.14 in terms of Cohen's kappa); thus the final dataset was annotated only with the emotions that the annotators perceived while reading;
- language: PO-EMO is the only multilingual dataset including German and English poems. Asian languages are represented by three datasets: one on Chinese (THU-FSPC) and two on Indo-Aryan languages (i.e. Punjabi in Kāvi and Odia in Kabithaa). The only dataset with texts written in a Classical language, i.e. Ancient Greek, is the one on Aeschylus, while the annotation of the *Iliad* is made of texts translated into Modern Greek.

Table 1
Datasets of poems annotated with sentiment information.

	ANNOTATORS	UNIT	CLASSES	PERSPECTIVE	LANG.
PO-EMO	experts/crowd	line/stanza	9	reader	DE/EN
DISCO PAL	experts	poem	7	author	ES
Kabithaa	experts	poem	2	author	OR
THU-FSPC	experts	poem/line	5	author	CN
Poem	experts	line	4	author	EN
PERC	crowd	poem	9	reader	EN
Kāvi	experts	poem	9	author	PA
Iliad	experts	line	4	reader	EL
Aeschylus	experts	sentence	5	reader	EL

Additional interesting annotated resources are those containing other textual genres based on versification, that is songs (Çano and Morisio 2017; Apoorva and Radhika 2018) and operas: as for the latter, AriEmozione 2.0 is made of lines taken from Italian 17th- and 18th-century operas (Zhang et al. 2022) annotated with one out of 6 emotions (love, joy, admiration, anger, sadness, fear).

Many of the datasets summarised in Table 1 have been used to train and test automatic systems using various types of supervised approaches and achieving very different results. Support Vector Machine (SVM) models reached an accuracy of 70% and 78% on Kāvi and Kabithaa respectively. Pre-trained BERT models have been instead tested on the German subset of PO-EMO and on the Poem corpus with an F1 measure of 52% in the former case and an accuracy of 85% in the latter. The first experiments on PERC adopted a Naive Bayes classifier (Sreeja and Mahalakshmi 2016) that achieved an accuracy of 35%. More recently, this dataset has been extended with new poems and a wider set of emotions (from 9 to 13); a BiLSTM classifier on this new version of PERC achieved an accuracy of 85% (Khattak et al. 2022).

With respect to the previous works presented in this section, in our annotation we chose to involve experts in Latin language and literature assigning four generic classes at the sentence level without defining the specific emotion conveyed by the text, and to focus on the sentiment as depicted by the author. Moreover, to account for the lack of training data for Latin poetry, in this paper we test a lexicon-based and a zero-shot classification approach.

3. Gold Standard Creation

This section provides details on the annotation procedure followed for creating the Gold Standard (GS).

3.1 Annotation

The GS consists of eight randomly selected odes of Horace, two from each of the four books that make up the work,[6] for a total of 955 tokens, without punctuation, and 44 sentences (average sentence length: 21, standard deviation: 11). Texts were taken from the corpus prepared by the LASLA laboratory in Liège.[7] We performed a single-label annotation of the original Latin text at sentence level. We have chosen the sentence as unit of annotation because it represents an intermediate degree of granularity between that of the verse and that of the stanza. In fact, the limited length of a verse can hinder the full understanding of the emotion polarity it conveys, while a stanza, being longer, risks to contain very different content and thus, potentially, even opposite polarities.[8] Furthermore, not all poems can be divided into stanzas, as this depends on the metric scheme of the poem. Instead, sentences can be detected in every poem regardless of their metric scheme, and represent a unit of meaning in their own right. The sentence division that we adopted is the one found in the LASLA corpus.

The annotation involved two experts in Latin language and literature (A1 and A2) and another annotator with basic knowledge of Latin but with previous experience in sentiment annotation (A3). Annotators were asked to identify the polarity conveyed by each sentence in the GS, taking into consideration both the vocabulary used by the author and the images that are evoked in the ode. More specifically, annotators were asked to answer the following question: which of the following classes best describes how are the emotions conveyed by the poet in the sentence under analysis?

- `positive`: the only emotions that are conveyed at lexical level and the only images that are evoked are positive, or positive emotions are clearly prevalent;
- `negative`: the only emotions that are conveyed at lexical level and the only images that are evoked are negative, or negative emotions are clearly prevalent;
- `neutral`: there are no emotions conveyed by the text;
- `mixed`: lexicon and evoked images produce opposite emotions; it is not possible to find a clearly prevailing emotion.

Sentences were presented to annotators in their original order in the poem, using a simple spreadsheet. To facilitate the recognition of emotion polarity and its assignment to the correct class, annotators adopted the Russell's Circumplex Model as a reference (Russell 1980): this model classifies a set of affect words according to their polarity

[6] Book I: odes 10 and 17; Book II: odes 7 and 13; Book III: odes 13 and 23; Book IV: odes 7 and 11.

[7] `http://web.philo.ulg.ac.be/lasla/opera-latina/`

[8] Note, however, that some sentences may even be longer than a stanza: for instance, the nr. 7 in our GS (*Od*. 1.17.21-8) covers in fact two full stanzas.

(whether it is positive or negative) and arousal (the intensity of the emotion). For example, *bored* has a low arousal and a negative polarity, while *excited* has a high arousal and a positive polarity. Even if the model is based on English words, they correspond to human emotions valid also for Latin and can help in identifying the correct label to assign.

The annotation of the GS was organised in various phases. At the beginning, annotators worked together collaboratively assigning the sentiment class to four of the eight odes (21 sentences): the task was discussed and a common procedure was defined. Later, annotators worked independently on the other four odes (23 sentences): A1 and A2 annotated the original Latin text, while A3 annotated the same odes using an Italian translation (Horace 2009) to understand whether the use of texts not in the original language can alter the annotation of emotion polarity. Then, we calculated the Inter-Annotator Agreement (IAA) and disagreements were discussed and reconciled.

3.2 Inter-Annotator Agreement

IAA was calculated on the 23 sentences independently annotated by all the three annotators (A1-A2-A3). The Fleiss's k among them resulted in 0.48, corresponding to what is considered a moderate agreement.[9] In particular, the `negative` class proved to be the easiest to annotate (Fleiss's k of 0.64), followed by `neutral` (0.57) and `positive` (0.45), whereas `mixed` was the most problematic class (0.23). Our IAA results are similar to those obtained during the annotation of `Poem` (overall Cohen's k of 0.50), that concerned the same textual genre (i.e., poetry) and the same four classes as our GS.

We noticed that the Italian translation was sometimes misleading, resulting in cases of disagreement: e.g., the sentence *inmortalia ne speres monet annus et almum quae rapit hora diem*, (ode IV, 7) is translated as 'speranze di eterno ti vietano gli anni e le ore che involano il giorno radioso' (literal translation of the Italian sentence into English: 'hopes of eternity forbid you the years and the hours that steal the radiant day'). A3 marked this sentence as `mixed`, considering that it is impossible to identify a prevailing emotion between the negativity expressed by the verb 'vietare' ('to forbid') and the positivity of 'giorno radioso' ('radiant day'). However, the translation of the Latin verb *rapio* is not appropriate to render the negative polarity of the original word: the Italian verb 'involare' ('to steal') does not fully convey the idea of the violent force inherent in *rapio*, which can be more correctly translated with the verb 'to plunder'.[10]

3.3 Consolidation

Disagreements were discussed and reconciled among the three annotators. Table 2 presents the number of sentences and tokens per polarity class. Our GS includes a majority of positive sentences (45.4%). Positive (average length: 21, standard deviation: 11), negative (average length: 24, standard deviation: 14), and mixed (average length: 25, standard deviation: 9) sentences are considerably longer than neutral ones (average length: 8, standard deviation: 3).

Four annotated examples are shown in Table 3 (English translations by Kaimowitz et al. (2008) are included for clarity), while Figure 1 shows how emotion polarities are

[9] For completeness of information, we report the IAA for combinations of the three annotators: the Cohen's kappa between A1 and A2 resulted in 0.5, between A1 and A3 was 0.38, between A2 and A3 was 0.56.

[10] See for instance the English translation by Kaimowitz et al. (2008): 'Do not hope for what's immortal, the year warns, and the hour which plunders the day'.

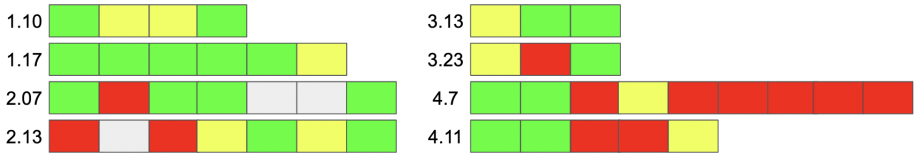

Figure 1
Pattern of emotion polarities in the sentences of the GS (green=`positive`; red=`negative`; yellow=`mixed`; grey=`neutral`).

distributed throughout the sentences in the GS. The distribution is different in each ode: some have a more positive orientation (i.e., 1.17), others show an evident alternation (i.e., 2.13). The fact that the sentences with negative polarity are mainly concentrated in the odes of the fourth book is not accidental. That book was written by Horace during his maturity and one of the main topic is the implacable passage of time, which leads to frequent melancholy or even pessimistic reflections (Porter 1975).

Table 2
GS statistics.

	Sentences	*Tokens*
`positive`	20	411
`negative`	12	292
`neutral`	3	23
`mixed`	9	229
TOTAL	44	955

Table 3
Annotated examples taken from the GS.

Ode	*Sent.*	*Text*	*Translation*	*Class*
1.17	103	*hic tibi copia manabit ad plenum benigno ruris honorum opulenta cornu*	Here for you will flow abundance from the horn that spills the country's splendors	`positive`
4.7	549	*cuncta manus auidas fugient heredis amico quae dederis animo*	All that you bestow upon your heart escapes the greedy hands of an heir	`negative`
2.13	265	*frigora mitescunt Zephyris uer proterit aestas interitura simul pomifer autumnus fruges effuderit et mox bruma recurrit iners*	With the Zephyrs cold grows mild, summer tramples springtime, soon to die, once productive autumn pours forth its fruits, and shortly lifeless winter is back	`mixed`
2.7	235	*quem Venus arbitrum dicet bibendi*	Who will Venus name as master of the wine?	`neutral`

4. Lexicon-Based Sentiment Analysis

This section provides details on the lexicon-based approach for the automatic analysis of our GS.

4.1 LatinAffectus

LatinAffectus is a manually-curated prior polarity lexicon currently comprising more than 5,000 Latin lemmas.[11] The entries of the lexicon are nouns and adjectives associated to numerical scores expressing their prior polarity, that is their sentiment orientation regardless of the context of use. Scores follow a five-value classification: -1.0 (fully negative, e.g. *inhonestus* 'shameful'), -0.5 (negative, e.g. *amaritudo* 'bitterness'), 0 (neutral, e.g. *cognatio* 'kinship'), +0.5 (positive, e.g. *uigens* 'active'), +1.0 (fully positive, e.g. *honos* 'dignity').

The development of *LatinAffectus* started in 2019 and is still ongoing: an expert annotator is periodically supported by university students with proven knowledge of the Latin language, with the aim of extending the lexicon as much as possible. The annotation follows the procedure defined in Sprugnoli et al. (2020b): first, annotators collaboratively assign polarity scores to a small set of lemmas in order to discuss the task, then they work independently using the aforementioned classification to which a class is added to indicate ambiguous lemmas, that is terms that cannot have a unique prior polarity due to their ambiguity. For example, *confidentia* means 'self-confidence' both in a good and in a bad sense depending on the context, thus it is not possible to assign it a positive or negative polarity. Later, the inter-annotator agreement (IAA) is calculated and discrepancies are discussed and reconciled with the help of a supervisor. Up to now, during the various development phases of *LatinAffectus* that involved different annotators, we recorded IAA results ranging from a minimum of 0.36 (fair agreement) to a maximum of 0.66 (moderate agreement) in terms of Cohen's k (macro average = 0.50).

Beside the entries to which polarity is manually assigned, *LatinAffectus* also contains a set of automatically created lemma-sentiment pairs. These additional pairs are derived starting from the manually-annotated entries by exploiting semantic and derivational relations and by adding graphical variants of lemmas. More specifically, a dictionary (Skřivan 1890) is used to find synonym and antonym (e.g. *pulcher* is a synonym of *formosus* 'beautiful' and *beneficium* 'benefaction' is an antonym of *maleficium* 'misdeed)', whereas morphological derivations are generated by 25 prefixal and suffixal relations taken from the database Word Formation Latin (Litta, Passarotti, and Culy 2016) (e.g. the prefix *in-* generates the adjective *incertus* 'uncertain' from *certus* 'determined' and the suffix *-(t)udo/udin* generates *laetitudo* 'joyful' from *laetus* 'joy').[12] Moreover, all the possible graphical variants of lemmas taken from the knowledge base of the LiLa project are added (e.g. *impurus* is a graphical variant of *inpurus* 'impure'). Original polarity scores are reversed for antonyms and for lemmas derived through the negative prefix *in-*, whereas they were preserved in all the other cases. This means that, starting from *certus* having a score of +0.5, *incertus* obtains a score of -0.5, whereas *certitudo* a score of +0.5.

[11] `https://github.com/CIRCSE/Latin_Sentiment_Lexicons/`

[12] Selected affixes are the following: *-ac(e/i), -al, -an, -ans/antis, -ar, -ari, -at, -bil, -e, -edo/edin, -ens/ent, -et, -i, -ic, -ici, -il, in (neg)-, -ist, -it, -iti, -ment, -n, -tas/tat, -(t)iu, -(t)udo/udin.*

The first release of *LatinAffectus* included 2,437 lemmas (Sprugnoli et al. 2020b) whereas the second release, called *LatinAffectus v2*, was made of additional 1,687 lemmas (Sprugnoli et al. 2021). For the purpose of the present work, we have added to the lexicon other 983 entries manually annotated and reconciled; thus, the version used in the paper, called *LatinAffectus v3*, has a total of 5,107 lemmas. We also employed an additional lexicon of more than 14,000 lemmas in which *LatinAffectus v3* is integrated with more than 9,000 lemmas annotated by only one annotator: this version is called *LatinAffectus v3+NR*, where NR stands for 'Not Reconciled' (see Table 4).

Table 4
Current composition of *LatinAffectus*.

	-1	-0.5	0	0.5	1	TOTAL
LatinAffectus v2	909 (22%)	542 (13,1%)	1,757 (42,7%)	461 (11,2%)	455 (11%)	4,124
LatinAffectus v3	969 (19%)	630 (12.3%)	2,524 (49.4%)	502 (9.8%)	482 (9.4%)	5,107
LatinAffectus v3 + NR	1,200 (8.5%)	1,470 (10.4%)	10,002 (71%)	760 (5.4%)	672 (4.8%)	14.124

4.2 Lexicon-based Polarity Classification

For this experiment we used the texts of Horace distributed as part of the LASLA corpus: the texts are manually annotated by experts with part of speech (PoS) tags, morphological features, and lemmas. The approach we adopted is a dictionary look-up of the lemmas present both in the Odes and in *LatinAffectus*. More specifically, the tokens in the *Odes* that are lemmatized under lemmas that also have an entry in *LatinAffectus* are assigned the score found in the lexicon. For instance, the adjective *malus* 'bad' has an assigned polarity value of -1.0 in *LatinAffectus*. All tokens lemmatized as *malus* (adj.) are thus given a score of -1.0. Note that a score of 0.0 is assigned to both words explicitly annotated as neutral in *LatinAffectus* and to those that do not have an entry in the lexicon.

The dictionary look-up required some disambiguation in cases of ambiguity due to homography. For 18 lemmas (corresponding to 49 tokens in the overall text of the *Odes*), the sentiment lexicon provides multiple values; in most cases, as with *ales* 'winged' (adj.), but also 'bird' (n.), the variation is due to a different polarity attributed to the syntactic uses of the word (in the example, to the adjective and the noun). In such cases, the PoS annotation in the LASLA corpus was used to disambiguate and match with the corresponding score in the lexicon. We also reviewed those words that, although not tagged as nouns or adjectives in LASLA, still yield a match in *LatinAffectus*. After revision, we decided to keep the scores for a series of lemmas annotated as numerals in the corpus (*simplex* 'simple, plain', *primus* and *primum* 'first', *prius* 'former, prior') and the indefinite pronoun *solus* 'alone, only' that in *LatinAffectus* are marked as adjectives.

A sentence score (S) was computed by summing the values of all lemmas. Thus, we attributed the label `positive` to all the sentences with score $S > 0$ and `negative` where $S < 0$. For $S = 0$, we attributed `neutral` to sentences where all words had a score of 0 and `mixed` where positive and negative scores were balancing each other out to a total net sum of 0. The overall accuracy of this method using *LatinAffectus v2* is 48%

(macro-average F1 37, weighted macro-average F1 44) with very diverse scores among the four classes: 70% for `positive`, 42% for `negative`, 67% `neutral`, while no correct predictions were given for `mixed`.

In order to check whether an increase in lexical coverage improves the scores of our first classification (even by including words that had not been reviewed by more than one annotator), we carried out two additional experiments. The first used *LatinAffectus v3* and the second *LatinAffectus v3+NR*.

Repeating the same steps with the two increased lexicons did not improve the performance in the classification task . On the contrary, the performances in terms of accuracy dropped from 48% to 45%. Interestingly enough, both the fully reviewed and the single-annotator versions of the dataset show an identical drop of the performances.

This lexicon-based approach completely overlooks the other syntactic and semantic aspects of the sentence. In particular, this method ignores the effect of negative particles that influence the context where a polarity-bearing word is found. Consider for instance the following sentence from the GS:

> *cum semel occideris et de te* ***splendida*** *Minos fecerit arbitria non Torquate genus non te* ***facundia*** *non te restituet* ***pietas***. 'When you at last have died and Minos renders **brillant** judgement on your life, no Torquatus, not birth, not **eloquence**, not your **devotion** will bring you back.' (*Od.* 4.7.21-4)

Here, the look-up on *LatinAffectus* yields a very positive score (3), which is caused by the highlighted, clearly positive words. However, the repeated negative particle (*non*) inverts the polarity of the positive words, and emphasises how the positive concepts 'eloquence' and 'devotion' are not sufficient to bring the dead back to life. Accordingly, the meaning of the sentence is very negative.

In order to establish the role of potential polarity shifters like negative particles, we introduced a simple modification to our classification system. In case one or more negative adverb was found, we reversed the overall polarity classification of the sentence (from positive to negative and vice versa). The negative particles that we selected were the following: *non* ('not'), *nec*, *neque*, *neue* ('and not', 'neither'), *ne* (negative conjunction), *numquam* ('never'). This method is clearly not sufficient to account for the effects of negative polarity shifters. Once again, instead of improving the scores of the simple look-up, the inversion of the classification causes a drop in accuracy for both sets: the overall accuracy for this method is 41% using *LatinAffectus v3* and 43% using *LatinAffectus v3 + NC*.

5. Zero-Shot Classification

We trained a language model for SA on English and tested it on our GS by relying on two state-of-the-art multilingual models. More specifically, we fine-tuned Multilingual BERT (mBERT) (Pires, Schlinger, and Garrette 2019) and XLM-RoBERTa (Conneau et al. 2020) with the GoEmotions corpus (Demszky et al. 2020) using the Hugging Face's PyTorch implementation.[13] GoEmotions is a dataset of comments posted on Reddit manually annotated with 27 emotion categories or Neutral. In order to adapt this dataset to our needs, we mapped the original emotions into coarse-grained polarity categories as suggested by the authors themselves. For example, joy and love were converged

[13] `https://huggingface.co/transformers/index.html`

into a unique `positive` class, whereas fear and grief were merged under the same `negative` class. The `neutral` category remained as is and comments annotated with emotions belonging to opposite sentiments were marked as `mixed`. Comments labeled with ambiguous emotions (i.e. realisation, surprise, curiosity, confusion) were instead left out.[14] With this procedure, we built a training set made of 18,617 positive, 10,133 negative, 1,965 neutral and 1,581 mixed comments, for a total of 32,296 texts. For fine-tuning, we chose the following hyperparameters: 32 for batch size, 2e-5 for learning rate, 6 epochs, AdamW optimizer.[15]

We evaluated the trained model on different datasets, including our GS. For each of them, we randomly selected 44 texts in order to have the same number of input data as in our GS:

- `GoEmotions`: test set taken from the same corpus used for training the English model;
- `Poem`: collection of English verses annotated with the same sentiment classes as in our GS, see Section 2;
- `AIT-2018`: English data released for the emotion classification task of "SemEval-2018 Task 1: Affect in Tweets" (Mohammad et al. 2018). Each tweet is annotated as neutral or as one, or more, of eleven emotions. The original categories were mapped onto our four sentiment classes, leaving out ambiguous emotions;
- `AriEmozione 2.0`: lines taken from a set of Italian opera texts annotated with one out of six emotions (Zhang et al. 2022). We mapped the original emotions to either the `positive` class or the `negative` class. Classes `mixed` and `neutral` are not present;
- `MultiEmotions-It`: a multi-labeled emotion dataset made of Italian comments posted on YouTube and Facebook (Sprugnoli 2020). The original emotion labels were converted into our four classes on the basis of their polarity.

The distribution of the four classes in the datasets used for training and testing the models is displayed, expressed in percentage values, in Figure 5. Each dataset has a different and quite unbalanced distribution: for example, English training data feature a majority of texts annotated as `positive`, while very few are annotated as `mixed` or `neutral`. On the contrary, `neutral` is very frequent in the test sets taken from the `Poem` dataset.

Table 5 reports the results of mono-lingual and cross-lingual classification for the different test sets briefly described above and for the two pre-trained multilingual models. There is no clear prevalence of one model over the other: results vary greatly from one dataset to another. On the same language (thus without zero-shot transfer), we notice a drop in the performance for both mBERT and XLM-RoBERTa when moving from Reddit comments, that is the same type of text as the training data, to tweets, but even more so when they are evaluated on poems. As for the zero-shot classification, results on Italian YouTube and Facebook comments are better than the ones registered on English tweets, but accuracy drops when applied to opera verses. However, the

[14] For the full mapping, please see: `https://github.com/google-research/google-research/blob/master/goemotions/data/sentiment_mapping.json`.

[15] We adapted the following implementation: `https://gist.github.com/sayakmisra/b0cd67f406b4e4d5972f339eb20e64a5`.

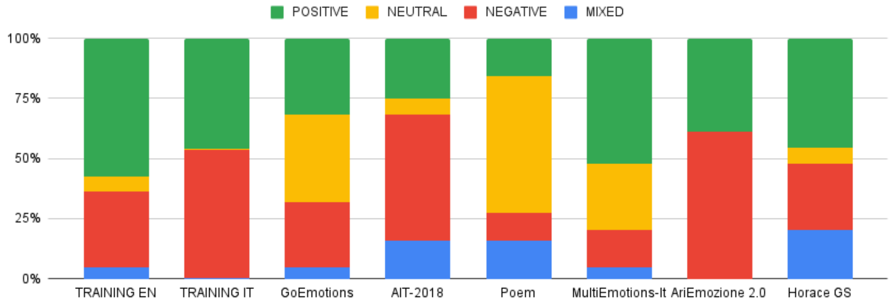

Figure 2
Distribution of classes in the datasets used as training and test sets.

worst results are recorded for Latin with an accuracy equal to, or slightly above 30% (for mBERT: macro-average F1 29, weighted macro-average F1 35; for XLM-RoBERTa: macro-average F1 24, weighted macro-average F1 26). For both mBERT and XLM-RoBERTa, we register the same trend at class level: perfect accuracy for `neutral`, good accuracy for `negative` (50% with mBERT and 67% with XLM-RoBERTa), low accuracy for `positive` (25% with mBERT and 10% with XLM-RoBERTa) and no correct predictions for `mixed`.

Table 5
Accuracy of the mono-lingual and cross-lingual (zero-shot) classification model fine-tuned on English.

Language	Test Set	Genre	mBERT	XLM-RoBERTa
English	`GoEmotions`	social network	86%	73%
	`AIT-2018`	social network	64%	59%
	`Poem`	literary - poetry	50%	70%
Italian	`MultiEmotions-It`	social network	70%	75%
	`AriEmozione 2.0`	literary - opera	59%	52%
Latin	`Horace GS`	literary - poetry	32%	30%

5.1 Impact of the Source Language

Given the low performance on Latin obtained by training the models on English, we carried out a further experiment by changing the source language. Indeed, the paper by de Vries et al. (2022) has shown that fine-tuning a multilingual model on English is not always the best option because some typological features, such as word order and lexical-phonetic distance, can impact cross-lingual performance. We thus repeated the zero-shot experiment using Italian (being a Romance language evolved from Latin) as source language instead of English. In order to have enough training data, we merged three natively Italian datasets annotated with emotions, i.e. `MultiEmotions-It`, `FEEL-IT` (Bianchi, Nozza, and Hovy 2021) and `AriEmozione 2.0`. The distribution of classes in this training set (TRAINING IT) is shown in the relevant column of Figure 5: the percentage of instances annotated as `neutral` and `mixed` is extremely low because these classes are not present in AriEmozione 2.0. Table 6 reports the new results of the models trained on Italian obtained on the same test sets used also in the previous experiment.

Table 6
Accuracy of the mono-lingual and cross-lingual (zero-shot) classification model fine-tuned on Italian.

Language	Test Set	Genre	mBERT	XLM-RoBERTa
English	GoEmotions	social network	52%	48%
	AIT-2018	social network	73%	68%
	Poem	literary - poetry	20%	27%
Italian	MultiEmotions-It	social network	75%	84%
	AriEmozione 2.0	literary - opera	82%	86%
Latin	Horace GS	literary - poetry	36%	48%

Comparing the accuracy scores in Tables 5 and 6, we can notice that the increase in the mono-lingual setting is greater with the XLM-RoBERTa model than with mBERT: with the former, the accuracy goes from 75% to 84% (+9 percentage points) on social network texts and from 52% to 86% (+34) on opera verses. As expected, the results on English are worse for both models: an improvement is recorded only on the AIT-2018 dataset. On our GS, XLM-RoBERTa performs better than mBERT with an improvement of 18 points with respect to the accuracy achieved using English as source language, going from 30% to 48% (for mBERT: macro-average F1 21, weighted macro-average F1 31; for XLM-RoBERTa: macro-average F1 26, weighted macro-average F1 40).

6. Discussion

Tables 7 and 8 report a comparison in terms of precision, recall and F1-score (both macro average and weighted average) between the lexicon-based approach and the zero-shot classification experiments run with both the mBERT and the XLM-RoBERTa models using Italian as training language. All the approaches achieve a better F1 on the positive class, and recall for this class is always higher than precision. The lexicon-based approach registered the highest F1 for the negative and the neutral classes but failed in the identification of the mixed class. Also the zero-shot model was not able to provide correct predictions for both the neutral and the mixed classes. Attempts to increase the percentage of instances labeled as mixed and neutral have not led to any performance improvements. In particular, we added to the Italian training data only the mixed and neutral instances of SENTIPOLC2016 (a collection of Italian Tweets labelled with subjectivity, polarity and irony) (Barbieri et al. 2016) and GoEmotions, the latter automatically translated into Italian with Google Translate, so to have a percentage of these two classes similar to that of the English training (i.e. between 5% and 6% of the total). Even so, the models were never able to assign either the mixed class or the neutral class, and overall accuracy even dropped (from 48% to 41% for XLM-RoBERTa).

A manual inspection of the output of the lexicon-based method reveals several shortcomings in that approach. The coverage of *LatinAffectus* has improved considerably throughout the stages of the present work: while the first set of ca 4,000 lemmas covered only the 46% of the nominal and adjectival lemmas in the *Odes*, the increased

Table 7
Macro and weighted average F1 for the lexicon-based method and for the zero-shot classification experiments trained on Italian.

F1	**Lexicon-Based**	**Zero-Shot mBERT**	**Zero-Shot XLM-RoBERTa**
macro avg	0.37	0.21	0.26
weighted avg	0.44	0.31	0.40

Table 8
Precision (P), recall (R) and F1-score (F1) for the lexicon-based method and for the zero-shot classification experiments trained on Italian.

	Lexicon-Based			**Zero-Shot mBERT**			**Zero-Shot XLM-RoBERTa**		
	P	R	F1	P	R	F1	P	R	F1
`positive`	0.56	0.70	0.62	0.45	0.50	0.48	0.53	0.80	0.64
`negative`	0.62	0.42	0.50	0.27	0.50	0.35	0.36	0.42	0.38
`neutral`	0.25	0.67	0.36	0.00	0.00	0.00	0.00	0.00	0.00
`mixed`	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

versions with the reconciled and the single-annotator lemmas now cover the 50% and the 84% of the tokens tagged as nouns or adjectives in the overall work. This improved coverage, however, does not entail an improved accuracy, as we saw. On the one hand, some lemmas with a clear sentiment orientation are still not captured, even with the single-annotator additions. For example, in the sentence:

> *Infernis neque enim tenebris Diana pudicum liberat Hippolytum nec Lethaea valet Theseus abrumpere caro vincula Pirithoo* 'For from the darkness below Diana does not free the chaste Hippolytus nor is Theseus able to break through the bonds that hod his Pirithous' (4.7.25-8)

only the tokens *pudicum* (lemma *pudicus* 'virtuous', score 0.5) and *caro* (lemma *carus* 'dear', score 1) are captured by the published version, while the single-annotator extension adds the negative score for the word *vinculum* (lemma of *vincula* 'chain'). The clearly negative noun *tenebrae* (lemma of *tenebris* 'darkness') is still not represented in the lexicon. Even if all the scores were added, however, positive and negative lexical scores would balance out and the sentence would be (incorrectly) classified as `mixed`, whereas the negative tone is undoubtedly dominating. The `mixed` class is, as we saw, virtually impossible to identify with the rules that we used. In order for a sentence to be classified as `mixed`, all the strongly and weakly positive/negative words must zero each other out; a slight deviation (e.g., a word that is slightly negative in the context of an *Ode* but classified as strongly negative in the lexicon) is sufficient to tilt the scale. Evidently, this level of exactitude does not reflect the judgment of the human readers who classified the sentences of the GS (which is, however, already controversial, as `mixed`, as we said, is the label with the lowest IAA score). In several cases of sentences manually labeled as `mixed` in the GS (e.g. *Odes* 1.17.21-8), the lexicon look-up correctly reflects the presence of both positive and negative words, even if the overall sum is not a perfect 0. On the other hand, the effect of the negative particles on the overall meaning of the sentence is still not adequately taken into account. In the example above, the positive word *caro* ('dear') is placed in a negated clause (*nec*... *valet*, 'has not... the strength'), which emphasises its pathetic effect ("dear as Pirithoos is to him, Theseus doesn't have the strength to take him back to life"). All these aspects cannot still be captured by a simple look-up on a polarity lexicon. On the other hand, a manual inspection of the results has confirmed to us that, although not very reliable as a method for SA, the dictionary look-up is an interesting exploratory technique, that is extremely useful to highlight trends and passages where a closer reading is particularly rewarding.

As for the zero-shot experiments, the problematic nature of Latin in this type of approach is confirmed by the fact that by processing the Italian translation of the Odes (Horace 2009) with the XLM-RoBERTa model (the best model in the mono-lingual setting) we achieved an accuracy of 52% (+12 points with respect to the accuracy registered with mBERT on the original Latin text and +10 points with respect to the accuracy achieved with XLM-RoBERTa). In particular, when applied to the translation, the model achieved a better accuracy on the `positive` class (which rises from 30% to 60%) and it was able to correctly classify the following sentence as `mixed`: 'Ultimo amore mio – non brucerò mai più per altre – impara a modulare con dolcezza i miei canti: leniranno la tua tristezza.' that is the Italian translation of *age iam meorum finis amorum non enim posthac alia calebo femina condisce modos amanda uoce quos reddas minuentur atrae carmine curae* (Od. 4.11.31-36).[16]

7. Conclusion and Future Work

In this paper, we have presented a number of experiments for the automatic classification of emotion polarity performed on the Odes of the Latin poet Horace. We defined new annotation guidelines suitable for capturing the complexity of the poetic text while maintaining a low number of labels. We built a small GS made of a random selection of eight odes of Horace and we evaluated two different approaches to automatic polarity classification that do not require Latin training data, namely: (1) a lexicon-based approach, based on a polarity lexicon of Latin (*LatinAffectus*), and (2) a zero-shot classification method. To the best of our knowledge, our GS and our experiments are the first to be carried out so far on Latin for the task of emotion polarity classification. Despite the limited size of our dataset and the non-positive results obtained in terms of classification performance, we think they are useful for paving the way for new annotations and new experiments.

The evaluation of the results provided by the two approaches shows that, although they are overall quite diverse, in terms both of the quantity and the distribution of wrong classifications, they both fail to correctly identify sentences with mixed sentiments, which, in any case, are the most problematic also for human annotators. The main limitations that impact the accuracy rates of the approaches that we described in this paper deal with the amount and the coverage of the available lexical and textual data supporting both the processing of input texts and the evaluation of the results produced. As for the zero-shot approach, the little number of available Latin corpora (in particular, Latin poems) enhanced with polarity annotation affects the reliability of the evaluation of the different automatic approaches that can be experimented. For the lexical-based approach, while the sheer size of the lexicon does not seem to have an impact on the classification, the lack of annotation on other levels of linguistic analysis,

[16] English translation: 'Come now, my last belovèd, — for I'll not be smitten by another woman after this— learn melodies from me to render with your lovely voice— with song black cares lessen.'

such as syntax, precludes the possibility to account for frequent phenomena of polarity inversion, like the role of negative particles that have polarity-bearing words in scope.

The very low performances of the zero-shot classification approach on Latin deserve further investigation. It is possible that the problem lies in the data used to build the pre-trained models: i.e., Wikipedia for mBERT and Commoncrawl for XLM-RoBERTa. Both resources were developed by relying on automatic language detection engines and are highly noisy due to the presence of languages other than Latin and of terms related to modern times.

Future work extends in three main directions. First, we want to expand the annotation to other Odes and to poems written by other Latin authors, possibly involving university students in order to analyse how IAA changes with annotators who did not participate in the discussion phase. Secondly, following the example of Vassallo et al. (2020), one strategy that can be tested in the future to improve the performances of the lexical-based approach consists in weighing the polarity scores of the *LatinAffectus* lemmas by their frequency in a reference corpus, then identifying the appropriate threshold for the three classes. The third line of research will focus on new experiments; for example, we can envisage the application of continued pretraining (Gururangan et al. 2020) or prompt-based learning (Liu et al. 2023). Moreover, we plan to test the feasibility of using LatinBERT, a model trained not only on the Latin Wikipedia but also on high-quality Latin corpora[17], as soon as more Latin annotated data will become available to fine-tune it.

Given the always growing interest in sentiment and emotion analysis and the challenges that the processing of Latin poses, we plan to add a task on emotion polarity classification in the next edition of EvaLatin, the campaign devoted to the evaluation of NLP tools for Latin, planned to be held in 2024. The first edition of EvaLatin, organized in 2020, focused on lemmatization and PoS tagging as shared tasks (Sprugnoli et al. 2020a). The second edition (2022) added one further task dedicated to the identification of morphological features (Sprugnoli et al. 2022). At the moment, we are building the data for two new tasks of the next edition of the campaign: i.e., dependency parsing and, as already said, polarity classification. These two tasks are indeed strictly connected. As a matter of fact, syntactically annotated textual data can prove much helpful for sentiment and emotion analysis purposes, especially when long sentences are concerned: the identification of the boundaries of the phrases where words that are assigned a polarity occur, as well as the recognition of the focus of negations might represent a substantial support to improve the accuracy of automatic systems.

In the long term, one of the biggest challenges in the area of sentiment and emotion analysis for Classical/Ancient languages consists of building sets of Latin texts enhanced with Word Sense Disambiguation annotation and developing automatic NLP tools. Indeed, while creating the GS described in this paper, we got through several occurrences of polysemous words, the automatic selection of whose specific contextual sense promises to improve systems performance.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

[17] https://github.com/dbamman/latin-bert

References

Apoorva, Gopikrishnan Drushti and Mamidi Radhika. 2018. Bolly: Annotation of sentiment polarity in bollywood lyrics dataset. In Kôiti Hasida and Win Pa Pa, editors, *Computational Linguistics*. Springer, Singapore, pages 41–50.

Baldo, Gianluigi. 2012. Horace (Quintus Horatius Flaccus), Carmina. In Christine Walde and Brigitte Egger, editors, *Brill's New Pauly Supplements I - Volume 5 : The Reception of Classical Literature*. Brill, Amsterdam.

Barbado, Alberto, Víctor Fresno, Ángeles Manjarrés Riesco, and Salvador Ros. 2022. DISCO PAL: Diachronic Spanish sonnet corpus with psychological and affective labels. *Language Resources and Evaluation*, 56(2):501–542.

Barbieri, Francesco, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*, pages 146–155, Napoli, Italy, December. Accademia University Press.

Bianchi, Federico, Debora Nozza, and Dirk Hovy. 2021. Feel-it: emotion and sentiment classification for the italian language. In *The 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 76–83, online, April. Association for Computational Linguistics.

Çano, Erion and Maurizio Morisio. 2017. Moodylyrics: a sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pages 118–124, Hong Kong, March. Association for Computing Machinery.

Chen, Huimin, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable chinese poetry generation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pages 4925–4931, Macao, August. International Joint Conferences on Artificial Intelligence Organization.

Conneau, Alexis, 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, July. Association for Computational Linguistics.

de Vries, Wietse, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from pos tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland, May.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 4040–4054, online, July. Association for Computational Linguistics.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, online, July. Association for Computational Linguistics.

Haider, Thomas, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. Po-emo: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1652–1663, online, May. European Language Resources Association (ELRA).

Herrmann, J. Berenike, Arthur M. Jacobs, and Andrew Piper. 2021. Computational stylistics. *Handbook of Empirical Literary Studies*, pages 451–486.

Horace. 2009. *I quattro libri delle Odi e l'Inno secolare di Quinto Orazio Flacco*. Flaccovio. Italian Translation by Gianfranco Nuzzo.

Kaimowitz, Jeffrey H., Ronnie Ancona, et al. 2008. *The odes of Horace*. Johns Hopkins University Press.

Khattak, Asad, Muhammad Zubair Asghar, Hassan Ali Khalid, and Hussain Ahmad. 2022. Emotion classification in poetry text using deep neural network. *Multimedia Tools and Applications*, 81:26223—26244.

Kim, Evgeny and Roman Klinger. 2018. A survey on sentiment and emotion analysis for computational literary studies. *arXiv preprint arXiv:1808.03137*.

Litta, Eleonora, Marco Passarotti, and Chris Culy. 2016. Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*, pages 185–189, Napoli, Italy, December. Accademia University Press.

Liu, Bing. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Mohammad, Saif M. 2012. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741.

Mohammad, Saif M., Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, pages 1–17, New Orleans, LA, USA, June. Association for Computational Linguistics.

Mohanty, Gaurav, Pruthwik Mishra, and Radhika Mamidi. 2018. Kabithaa: An annotated corpus of Odia poems with sentiment polarity information. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Passarotti, Marco, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Pavlopoulos, John, Alexandros Xenos, and Davide Picca. 2022. Sentiment Analysis of Homeric Text: The 1st Book of Iliad. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7071–7077, Marseille, France, June. European Language Resources Association (ELRA).

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Porter, David H. 1975. The recurrent motifs of Horace, Carmina IV. *Harvard Studies in Classical Philology*, 79:189–228.

Pozzi, Federico, Elisabetta Fersini, Enza Messina, and Bing Liu. 2016. *Sentiment analysis in social networks*. Morgan Kaufmann.

Russell, James A. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Saini, Jatinderkumar R. and Jasleen Kaur. 2020. Kāvi: An annotated corpus of punjabi poetry with emotion detection based on 'navrasa'. *Procedia Computer Science*, 167:1220–1229.

Schmidt, Thomas and Manuel Burghardt. 2018. An Evaluation of Lexicon-based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 139–149, Santa Fe, NM, August. Association for Computational Linguistics.

Sheng, Emily and David C. Uthus. 2020. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106, online, December. Association for Computational Linguistics.

Skřivan, Arnošt. 1890. *Latinská synonymika pro školu i dum*. V CHRUDIMI.

Sprugnoli, Rachele. 2020. MultiEmotions-it: A new dataset for opinion polarity and emotion analysis for Italian. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 402–408, online, March. Accademia University Press.

Sprugnoli, Rachele, Francesco Mambrini, Giovanni Moretti, and Marco Passarotti. 2020. Towards the Modeling of Polarity in a Latin Knowledge Base. In *Proceedings of the Third Workshop on Humanities in the Semantic Web (WHiSe 2020)*, pages 59–70, online, June. CEUR Workshop Proceedings (CEUR-WS. org).

Sprugnoli, Rachele and Marco Passarotti, editors. 2020. *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, France, May. European Language Resources Association (ELRA).

Sprugnoli, Rachele, Marco Passarotti, Flavio Massimiliano Cecchini, and Matteo Pellegrini. 2020a. Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France, May. European Language Resources Association (ELRA).

Sprugnoli, Rachele, Marco Passarotti, Daniela Corbetta, and Andrea Peverelli. 2020b. Odi et Amo. Creating, Evaluating and Extending Sentiment Lexicons for Latin. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3078–3086, online, May. European Language Resources Association (ELRA).

Sprugnoli, Rachele, Marco Passarotti, Cecchini Flavio Massimiliano, Margherita Fantoli, and Giovanni Moretti. 2022. Overview of the evalatin 2022 evaluation campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), Language Resources and Evaluation Conference (LREC 2022)*, pages 183–188, Marseille, France, June. European Language Resources Association (ELRA).

Sprugnoli, Rachele, Marco Passarotti, Marinella Testori, and Giovanni Moretti. 2021. Extending and using a sentiment lexicon for latin in a linked data framework. In *Workshop on Sentiment Analysis and Linguistic Linked Data (SALLD-1)*, pages 1–14, Zaragoza, Spain, September. CEUR Workshop Proceedings (CEUR-WS. org).

Sreeja, Ponnarassery Sreenivasan and GS Mahalakshmi. 2016. Emotion recognition from poems by maximum posterior probability. *International Journal of Computer Science and Information Security*, 14:36–43.

Vassallo, Marco, Giuliano Gabrieli, Valerio Basile, and Cristina Bosco. 2020. Polarity Imbalance in Lexicon-based Sentiment Analysis. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*, pages 457–463, online, March. Accademia University Press.

Volkova, Ekaterina P., Betty Mohler, Detmar Meurers, Dale Gerdemann, and Heinrich H Bülthoff. 2010. Emotional perception of fairy tales: achieving agreement in emotion annotation of text. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 98–106, Los Angeles, CA, June.

Yeruva, Vijaya Kumari, Mayanka ChandraShekar, Yugyung Lee, Jeff Rydberg-Cox, Virginia Blanton, and Nathan A. Oyler. 2020. Interpretation of sentiment analysis in aeschylus's greek tragedy. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 138–146, online, December.

Zehe, Albin, Martin Becker, Lena Hettinger, Andreas Hotho, Isabella Reger, and Fotis Jannidis. 2016. Prediction of happy endings in german novels based on sentiment information. In *3rd Workshop on Interactions between Data Mining and Natural Language Processing*, pages 9–16, Riva del Garda, Italy, September.

Zhang, Shibingfeng, Francesco Fernicola, Federico Garcea, Paolo Bonora, and Alberto Barrón-Cedeño. 2022. Ariemozione 2.0: Identifying emotions in opera verses and arias. *Italian Journal of Computational Linguistics (IJCoL)*, 8(8-2).