Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3078–3086
Marseille, 11–16 May 2020

Odi et Amo
Creating, Evaluating and Extending Sentiment Lexicons for Latin

Rachele Sprugnoli, Marco Passarotti, Daniela Corbetta, Andrea Peverelli
CIRCSE Research Centre, Università Cattolica del Sacro Cuore
Largo Agostino Gemelli 1, 20123 Milano
{rachele.sprugnoli,marco.passarotti}@unicatt.it

Abstract

Sentiment lexicons are essential for developing automatic sentiment analysis systems, but the resources currently available mostly cover modern languages. Lexicons for ancient languages are few and not evaluated with high-quality gold standards. However, the study of attitudes and emotions in ancient texts is a growing field of research which poses specific issues (e.g., lack of native speakers, limited amount of data, unusual textual genres for the sentiment analysis task, such as philosophical or documentary texts) and can have an impact on the work of scholars coming from several disciplines besides computational linguistics, e.g. historians and philologists. The work presented in this paper aims at providing the research community with a set of sentiment lexicons built by taking advantage of manually-curated resources belonging to the long tradition of Latin corpora and lexicons creation. Our interdisciplinary approach led us to release: i) two automatically generated sentiment lexicons; ii) a Gold Standard developed by two Latin language and culture experts; iii) a Silver Standard in which semantic and derivational relations are exploited so to extend the list of lexical items of the Gold Standard. In addition, the evaluation procedure is described together with a first application of the lexicons to a Latin tragedy.

Keywords: sentiment analysis, Latin, evaluation.

1. Introduction

The most common applications of resources and tools for sentiment analysis, i.e. the task of automatically classifying a piece of text according to the sentiment conveyed by it, fall into categories like social media and customer experience monitoring, or people's opinion mining. Accordingly, the texts which are most frequently analysed with sentiment analysis tools, and included in linguistic resources for sentiment analysis, are the kind of tweets, comments, feed-backs and social media chats. These are also the most widespread textual typologies used for building such tools in data-driven fashion, by exploiting a wide set of empirical data for training.

Said characteristics of sentiment analysis tools and resources (i.e. applications, textual typologies and size of data) do not take place when historical and ancient languages are concerned. If available, the most frequent application of sentiment analysis tools and resources in such area would deal with evaluating the lexical properties of literary, philosophical, or documentary texts, to assess their degree of positive, neutral, or negative sentiment. Unlike social media texts, which grow hourly in size, texts written in historical and ancient languages are provided by closed corpora, which very rarely can be extended thanks to newly found texts. The Open Greek and Latin project[1], whose ultimate goal is to represent every source text produced in Classical Greek or Latin during antiquity (up to c. 600 AD) with the outlook of covering also the post-Classical era until present times, calculated that around 150 million words of Ancient Greek and Latin survive from antiquity, while more than 200 million words of post-Classical Latin result from the analysis of 10,000 books downloaded from Archive.org. While such numbers may look quite impressive to a Classicist, actually they are not if we consider that the most recent deep/machine learning techniques make use of billions of data .

Despite such quantitative limitation, the research area dedicated to building and using linguistic resources for ancient and historical languages has seen a substantial growth during the last decade. This has primarily concerned Latin and Ancient Greek as essential media for accessing and understanding the so-called Classical tradition.

Although Latin was among the first languages to be automatically processed with computers, thanks to the pioneering work on the texts of Thomas Aquinas by the Italian Jesuit Roberto Busa in the 40s (Busa, 1974), throughout its 60-year long history, computational linguistics has been mainly focusing on living languages, because of their larger economic and social impact compared to Classical/dead ones. However, the start, in 2006, of two independent (but related) projects aimed at building the first treebanks for Latin gave rise to a kind of renaissance for the research area of linguistic resources and NLP tools for ancient languages (Bamman et al., 2008). The results of such research area promise to impact a large and diverse community made of historians, philologists, archaeologists and literary scholars, in different ways all dealing with textual and lexical data in Latin.

This research is performed in the context of the *LiLa: Linking Latin project* (2018-2023)[2] (Passarotti et al., 2019) which aims at building a Knowledge Base of linguistic resources for Latin based on the Linked Data paradigm, i.e. a collection of several data sets described using the same vocabulary of knowledge description and linked together. Within the LiLa project, aside from interlinking the already available resources for Latin, we are also building a number of new ones, among which is an extended and checked version of the Latin WordNet (Franzini et al., 2019) and a

[1] http://www.opengreekandlatin.org/

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

set of Latin Sentiment Lexicons, which assign a sentiment score to a basic set of Latin adjectives and nouns. At first, we decided to focus on adjectives and nouns only, because their sentiment score seems to be more easy to define at a lexical level, i.e. out of context, than that of verbs. Indeed, the semantics of verbs is more strictly connected to that of the lexical items filling their argument positions. According to the basic statement of Fillmore's Frame Semantics, the meaning of some words can be fully understood only by knowing the frame elements that are evoked by those words (Fillmore, 1982), which is particularly relevant for valency-capable words, the largest number of which are verbs.

The fact of having focused on a first set of parts of speech whose sentiment score seemed (at least in principle) easier than another to assign to, is strictly connected to the fact that building such resources for Latin is made complex by a number of socio-linguistic properties of Latin, which may impact its lexical semantic shift. Such properties include the wide diachrony and diatopy of Latin texts, spread all over Europe and the Mediterranean area through a period of more than two millennia, and the absence of native speakers, which is a remarkable issue when dealing with the assignment of sentiment scores to lexical items.

This paper describes the building of a set of Latin Sentiment Lexicons whose contents are planned to be included in the LiLa Knowledge Base. In particular, the paper details the different methodologies we tested to create the lexicons and the evaluation process we designed. Moreover, the creation of a manually-curated Gold Standard is presented together with its extension generated by exploiting a rich set of linguistic information taken from available digital resources for Latin.

2. Related Work

In the last decade, the attention towards sentiment analysis and related tasks, such as opinion mining and emotion analysis, has significantly increased in both the academic and the business fields (Pang et al., 2008; Liu, 2015). Automatic systems have been developed to assign a positive, negative or neutral label to texts of different kinds, in particular reviews of products and services (Fang and Zhan, 2015) and social media posts (Nakov et al., 2016). Sentiment lexicons are essential resources for the development of such systems: they are structured as lists of words (and, in some cases, also phrases and multi-word expressions) associated to scores expressing their prior polarity, that is their sentiment orientation regardless of the context of use.

Since the manual creation of sentiment lexicons is a very time-consuming process, several automatic techniques have been developed to reduce the annotation effort, among which are cross-lingual projection methods and induction methods based on dictionaries, corpora and word embeddings. As for the former, bilingual resources, such as dictionaries and parallel corpora, are used to translate a lexicon from one language to another (Mihalcea et al., 2007). As for the latter, dictionary-based methods induce a list of polar terms from lexical databases or monolingual thesauri whereas corpus- and word embeddings-based methods operate on raw texts and vector representations of words respectively. All induction methods rely on a set of seed terms, that is a manually curated list of polar terms with a clear sentiment orientation. Starting from the work by Hu and Liu (2004), many works apply bootstrapping algorithms to the English WordNet (Miller, 1995) and analogous resources available for other languages (Sidarenka and Stede, 2016). Co-occurrence patterns and label-propagation algorithms are instead adopted for the automatic induction from unlabeled texts (Takamura et al., 2005; Velikovich et al., 2010). The label-propagation approach is also used in combination with word embeddings for the creation of domain-specific sentiment lexicons (Hamilton et al., 2016).

To the best of our knowledge only two sentiment lexicons for Latin have been released so far, but without a thoughtful evaluation of their quality. The first one was generated with the automatic translation of the NRC VAD lexicon, a resource created through crowdsourcing annotations (Mohammad, 2018), whereas the second was produced with a knowledge graph propagation algorithm starting from Wikipedia (Chen and Skiena, 2014). By processing the entries of these two lexicons with the Latin morphological analyzer and lemmatizer LEMLAT v3[3] (Passarotti et al., 2017) and by checking the results manually, it resulted that they both contain noisy data, that is lemmas from languages different than Latin, e.g. 'aaaaaaah' and 'aforementioned': 14% in the NRC VAD lexicon and 9% in the other.

Another approach is given by the API of the *Latin WordNet* project of the University of Exeter which can perform sentiment analysis of individual strings via HTTP POST requests[4] because each Latin synset incorporates the sentiment scores provided by the English SentiWordNet (Baccianella et al., 2010). This service has two main drawbacks: (a) it does not include a word sense disambiguation system and (b) it is based on a resource which contains modern senses inherited from the English language. For instance, the adjective *incompatibilis* 'incompatible' inherits a sense related to computers and defined as '(computers) incapable of being used with or connected to other devices or components without modification'.

Differently from the previously mentioned works, our study aims at: i) providing the community of scholars in both NLP and Classical studies with new high-quality resources; ii) applying and evaluating well-known methodologies to a dead language; iii) exploiting a set of already available manually curated linguistic resources to create, evaluate and extend our sentiment lexicons for Latin.

3. Methods

In this Section we describe the two approaches we followed to automatically generate Latin sentiment lexicons. The first one is based on the cross-lingual projection of sentiment scores using bilingual lexicographic resources; the second operates on the distributed vector representations of lemmas.

[3] https://github.com/CIRCSE/LEMLAT3

[4] https://latinwordnet.exeter.ac.uk/sentiment/

3.1. Cross-Lingual Projection of an English Lexicon

The first method we adopted takes advantage of bilingual dictionaries to translate the entries of an English sentiment lexicon into Latin. More specifically, two lexicographic resources available in digital format were used as bridges between English and Latin, allowing to map the sentiment scores registered in the lexicon of the reference language (i.e. English) to a new lexicon in the target language (i.e. Latin).

As for the English lexicon, we relied on the resource created by Cho et al. (2014) that merges 10 sentiment lexicons (e.g., General Inquirer (Stone and Hunt, 1963) and SentiWordNet) by standardizing and averaging the different values at the entry word level. The output of this process is a sentiment lexicon of 26,193 entries associated to fine-grained scores between -1 (fully negative, e.g. *abominable*) and +1 (fully positive, e.g. *breathtaking*).

We extracted the Latin entries and their English translations from the William Whitaker's Words digital dictionary[5] and the Cassell's Latin dictionary[6] (Simpson, 1959) in order to obtain a bilingual lexicon of 24,623 adjectives and nouns. Not surprisingly, the relation between Latin lemmas and English translations is not 1:1; a Latin lemma can have more than one translation (e.g., the adjective *genuinus* is translated both as 'natural' and 'relating') and, vice versa, an English translation can be associated to more than one Latin lemma (e.g. 'beautiful' is associate to both *pulcher* and *speciosus*).

The cross-lingual projection of the original score to the Latin translation of English lemmas resulted in a sentiment lexicon of 10,516 Latin lemmas. For example, the English term 'crime' was translated into 17 different Latin nouns such as *noxa* and *nefarium* all inheriting the score -0.741 originally associated to the English term.

3.2. Induction with NWE-Based Methods

The second method we tested is based on a set of algorithms that induce sentiment lexicons from neural word embeddings (NWE) starting from a list of seed terms with known sentiment scores. In particular, we adopted the algorithms implemented by Sidarenka (2019) provided by the Sentiment Lexicon Generation Suite (SentiLex)[7].

To create our list of seed terms we extracted the 200 most frequent adjectives and nouns appearing in "Opera Latina" (Denooz, 2004), a corpus of 158 texts written by 20 Classical authors, all manually annotated with lemmas and Part-of-Speech (PoS) tags. Two experts of Latin language and culture collaboratively assigned a sentiment score to these lemmas using a five-value classification: 1 (fully positive), 0.5 (somewhat positive), 0 (neutral), -0.5 (somewhat negative), -1 (fully negative). At the end, 129 terms with a clear, unambiguous sentiment score (fully positive, neutral and fully negative) were used as seeds. The other 71 lemmas were left out of the seed term lists for two possible reasons: i) 56 were not included because they had an intermediate score (e.g, 0.5 or -0.5), ii) 15 because they were ambiguous (e.g. *fortuna* can mean both 'good luck' than 'bad luck'). The embeddings employed for our experiments were pre-trained on the "Opera Latina" corpus as well: manual annotations were used to convert each token into a `LEMMA_PoS` representation so to preserve the information about the grammatical category of each lemma.

We used word2vec representation (Mikolov et al., 2013), 100 dimensions and the Continuous Bag-of-Words (CBOW) model (Sprugnoli et al., 2019). All the three algorithms available in the SentiLex release have been tested: (i) nearest centroids, (ii) k-NN clustering, (iii) principal component analysis (PCA). As suggested by Sidarenka (2019), mean scaling and length normalization were performed on input vectors before passing them to the algorithms. The output of each algorithm is a lexicon of polar (i.e. either positive or negative) terms.

4. Evaluation

In order to evaluate the quality of the sentiment lexicons automatically generated with the approaches described in the previous Section, we created a Gold Standard (GS) through a multi-stage process. We then measured the accuracy of the scores in the lexicons compared to the manually-assigned scores.

4.1. Creation of the Gold Standard

The GS was built by randomly selecting 1,040 lemmas from the lexicon generated with the cross-lingual projection method. The distribution of adjectives and nouns (35% versus 65%) follows the same statistical distribution of PoS in the lexicon.

Score assignment In the first phase of the annotation, two Latin experts collaboratively assigned sentiment scores to 20 adjectives and 20 nouns in order to discuss the task and define a common procedure. For example, in this phase the annotators chose the reference dictionaries to consult in case of doubt about the meanings of lemmas: first the Oxford Latin Dictionary (Souter, 1968) and then, if supplementary information had been necessary, the Vocabolario della Lingua Latina (Castiglioni and Mariotti, 1996). In the second phase, the experts independently assigned a sentiment score to each lemma using a six-value classification[8]: 1 (fully positive), 0.5 (somewhat positive), 0 (neutral), -0.5 (somewhat negative), -1 (fully negative), 2 (ambiguous). This last class marks terms that have a semantic and/or diachronic ambiguity and because of this ambiguity they cannot have a unique a priori score. For example, *pusillus* has a semantic ambiguity: it literally means 'somewhat very little' having thus a neutral meaning, but it can also be used in a very negative moral sense to address wretched people, despicable and immoral members of society. An example of diachronic ambiguity is *regalis* 'regal', whose meaning varies heavily throughout the ages. Starting from the very first centuries of Roman history, being it a kingdom, Latin speakers viewed it as a neutral or

[5] `https://mk270.github.io/whitakers-words/`

[6] This dictionary is available online in XML Dictionary eXchange format: `https://github.com/nikita-moor/latin-dictionary`

[7] `https://github.com/WladimirSidorenko/SentiLex`

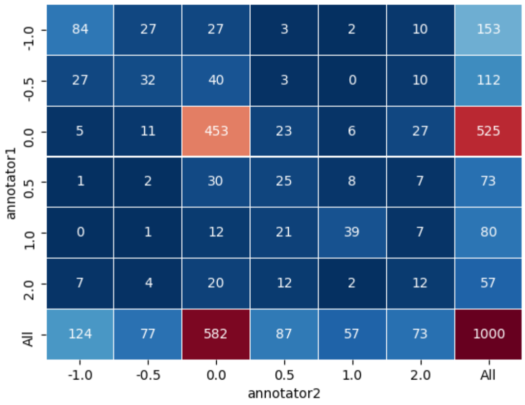
[8] The annotation process required a total of 20 person hours.

even positive term. Then it shifted to negative during the Republican Age. After the beginning of the Imperial Age, the term, now being related to Emperors, began again to be viewed as positive. It was also inherited by the Medieval Age, during which it pertained to all things regal and divine. Finally, it passed on to the temporal power of the Church, where it maintained its strong positive meaning, and it began to be referred to popes, cardinals and bishops.

Inter-Annotator Agreement After this second phase, we calculated the Inter-Annotator Agreement (IAA) on 1,000 lemmas, thus not taking into consideration the 40 terms used in the first phase. The Cohen's *k* was measured both on the six-value classification and also on a four-value classification in which the score 1 and 0.5 were merged into a unique positive class, whereas -0.5 and -1 were converged under the same negative class. As reported in Table 1, the agreement resulted as moderate (Artstein and Poesio, 2008): reducing the number of classes from 6 to 4 increases the agreement of 0.10 points. Nouns proved to be easier to be annotated in a consistent way.
Figure 1 shows the confusion matrix for the six-value classification: scores that tend to be confused between each other can be identified by noting non-diagonal cells with high values. Analysing the figure, we can notice that: i) the class of neutral terms (score=0) was the most frequent, ii) the distinction between neutral and somewhat positive terms was not always clear-cut and iii) ambiguity (score=2) was often confused with neutrality.

	ADJ	NOUN	MACRO-AVG
6 CLASSES	0.39	0.49	0.45
4 CLASSES	0.49	0.59	0.56

Table 1: Results of the IAA in terms of Cohen's *k* before the reconciliation.



annotator1 \ annotator2	-1.0	-0.5	0.0	0.5	1.0	2.0	All
-1.0	84	27	27	3	2	10	153
-0.5	27	32	40	3	0	10	112
0.0	5	11	453	23	6	27	525
0.5	1	2	30	25	8	7	73
1.0	0	1	12	21	39	7	80
2.0	7	4	20	12	2	12	57
All	124	77	582	87	57	73	1000

Figure 1: Confusion matrix among annotators on score assignment to both adjectives and nouns using a six-value classification. Value 2 was assigned to ambiguous lemmas.

Reconciliation In the last phase, the two Latin experts met to analyse the discrepancies in their annotations and reconcile all the cases of disagreement. This process led to the removal of 81 lemmas from the original list of 1,000 terms because of their ambiguity.

SCORE	ADJ	NOUN
1	*comis* 'gracious'	*honos* 'dignity'
0.5	*uigens* 'active'	*magister* 'master'
0	*arenosus* 'sandy'	*buculus* 'steer'
-0.5	*hebes* 'stupid'	*amaritudo* 'bitterness'
-1	*inhonestus* 'shameful'	*noxia* 'crime'

Table 2: Examples taken from the Gold Standard.

It is worth noticing that the assignment of scores required not only lexical expertise but also a profound knowledge of Latin political, social and religious culture. For example, the noun *monstrum*, despite the meaning of its corresponding terms in modern languages (such as 'mostro' in Italian), is associated to the concept of 'prodigy', but in a supernatural sense: it was the incarnation of a deity among humans, a god that showed him/herself (hence *monstrum* < *monstrare* = 'to show') through natural portents. It also has a very intense meaning of awe, and somewhat fear, inspiring respect and submission but in a generally positive way, because of its divine origin. For this reason, the score of *monstrum* is 0.5.
The final GS was then assembled by adding to the list of 919 unambiguous lemmas (that is, 1,000 initial lemmas - 81 ambiguous lemmas) the 129 terms used as seeds in the induction method, the 56 lemmas with intermediate scores taken from 'Opera Latina' and the 40 terms collaboratively annotated in the first phase so to obtain a final list of 1,144 adjectives and nouns. Examples taken from the GS are displayed in Table 2 together with their score. Note that, after removing ambiguous lemmas, no term with score = 2 was registered in the GS thus obtaining a final classification with 5 scores, i.e. 1, 0.5, 0, -0.5, -1.

4.2. Accuracy of Generated Sentiment Lexicons

Once built the GS, we calculated the accuracy of automatically generated sentiment lexicons. Table 3 reports the results of this evaluation taking into account both adopted methods.
As for the translation approach, we evaluated the lexicon after converting the original fine-grained scores to the same five-value classification of the GS. We also evaluated the same lexicon considering a more coarse-grained classification with 3 scores: 1 (positive), 0 (neutral), -1 (negative). Results show that the granularity of the scores has a great impact on the accuracy: removing the often subtle distinction between 0.5 and 1 and between -0.5 and -1 increases the micro-average accuracy of more than 15 points.
As for the induction method, the k-NN algorithm performed remarkably better than nearest centroids and principal component analysis algorithms (+48.5 and +47.9 respectively) whose results are definitely lower than the ones obtained with the cross-lingual projection method (their accuracy is below 30%).
In the lexicon built with the cross-lingual projection method, nouns have a higher accuracy than adjectives; the contrary is noted instead in lexicons produced with the induction method. The output of k-NN algorithms is particularly skewed (the difference in terms of accuracy be-

tween nouns and adjectives is 24.2 points) whereas results obtained with the cross-lingual projection method are more balanced.

	PROJECTION		INDUCTION		
	5 CL	3 CL	NC	k-NN	PCA
Adj	44.3%	64.9%	31.8%	86.7%	32.1%
Noun	54.8%	66.8%	21.7%	62.5%	22.3%
Micro-Avg	50.61%	66.1%	25.9%	**74.4%**	26.5%

Table 3: Accuracy of the automatically generated lexicons compared to the Gold Standard.

The k-NN algorithm calculates the distance between the vectors of seed terms and the vector of a lemma *l* and then assigns to *l* the sentiment score of the seed that is closest to *l* and appears most often as *l*'s neighbor. The resulting lexicon is thus a list of lemma-sentiment pairs ranked in ascending order according to the distance between vectors. Table 4 shows the top-scoring lemmas with their corresponding sentiment as generated by this algorithm: we can notice that they are exclusively true polar terms and that the sentiment was correctly assigned.

Lemma	PoS	Sentiment
miseria 'misery'	noun	negative
cruciatus 'torture'	noun	negative
optabilis 'desiderable'	adj	positive
beneuolentia 'good-will'	noun	positive
aerumna 'trouble'	noun	negative

Table 4: Top-5 polar terms produced by the k-NN algorithm.

As an additional evaluation, we calculated the accuracy of the lexicon of positive and negative Latin words generated through the knowledge graph propagation method as released by Chen and Skiena (2014) using the same GS. The obtained accuracy of 62.1% is lower than the one registered on the lexicons generated with the cross-lingual projection method (3 classes) and the k-NN algorithm (-4 and -12.3 points respectively).

5. Gold Standard Extension

Together with the GS, we released a Silver Standard built by extending the manually annotated list of lemma-sentiment pairs with other pairs obtained by exploiting different types of linguistic relations and taking advantage of 3 resources for Latin:

1. the dictionary of Latin synonyms compiled by Skřivan (1890), and available online in XML format[9], which allowed us to derive new sentiment-related lemmas through synonym and antonym relations with known lemmas in the GS (e.g., *pulcher* 'beautiful' → *formosus* 'handsome'; *beneficium* 'favor' → *maleficium* 'offence').

2. the Word Formation Latin[10] (WFL) database (Litta et al., 2016), a derivational morphology resource made of lemmas, analysed terms of input/output relations. Relations between lemmas are based on word formation rules: we selected a set of 25 prefixal and suffixal relations[11] and expanded the GS through bidirectional morphological derivations generated by such relations (e.g. for suffix *-(t)udo/udin* we had the expansion *laetus* 'joyful' → *laetitudo* 'joy' and also *amaritudo* 'bitterness' → *amarus* 'bitter'). We chose the aforementioned affixes because their effect on the original sentiment score is predictable and not ambiguous. On the contrary, we decided not to include affixes that can have different effects on the basis of the context of use: for example, *-ul* is used as diminutive but also as a meliorative or pejorative. Indeed, the noun *amiculus* can mean both 'dear friend' or 'humble friend'.

3. the list of all the possible written representations of the same lemma as available in the knowledge base of the *LiLa: Linking Latin* project. In Lila, Latin linguistic resources are connected to each other following the principles of the Linked Data framework. In this context, lemmas are used as a key node in the network of linguistic information where changes in spelling (e.g. *improsper* - *inprosper* 'unfortunate') and ending (e.g. *tropaeom* - *tropaeum* 'trophy') are managed as different written representations of the same lemma. This resource allowed us to add graphical variants of the lemmas already in the GS, instead of adding new lemmas.

	#	SCORE
Synonyms	727	propagating
Antonyms	123	reversing
Morphology: in(neg)-	373	reversing
Morphology: suffixes	5,254	propagating
Written Rep	25,923	propagating

Table 5: Type and number of linguistic relations used for Silver Standard creation and their effect on the score of the original lemma present in the Gold Standard.

Following (Neviarouskaya et al., 2009), the aforementioned linguistic relations were classified into two main groups on the basis of the effect they have on the sentiment score of the original lemma in the GS (see Table 5 for details): i) propagating, the sentiment score of the original lemma is preserved and propagated to the newly derived lemma, and ii) reversing, the sentiment score of the original lemma becomes the opposite when assigned to the newly derived lemma. These propagating and reversing relations have been applied recursively so to produce derivational chains. In Figure 2 an example of extension is visualized: starting from the lemma *purus* 'pure' included in the GS, we were able to add other 8 entries to the Silver Standard by using

[9] https://github.com/nikita-moor/latin-dictionary/tree/master/Skrivan1890

[10] http://wfl.marginalia.it/

[11] 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*.

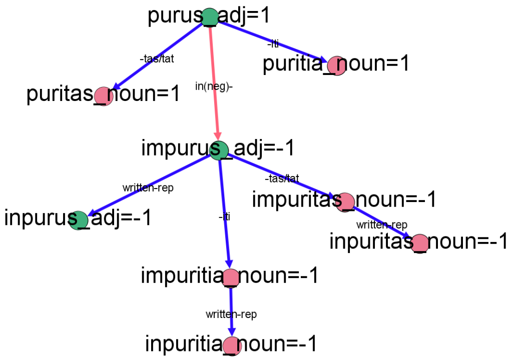


Figure 2: Example of extension from the lemma *purus* 'pure' present in the Gold Standard. Node color represents the PoS (green for adjectives, pink for nouns), edge color discriminates between propagating (in blue) and reversing (in red) relations with respect to the sentiment score.

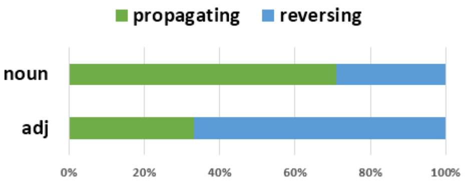


Figure 3: Percentage distribution of adjectives and nouns derived through propagating and reversing relations.

derivational relations (i.e. with affixes *in (neg)-*, *-tas/tat* and *-iti*) and different written representations. The *in (neg)-* prefix produced an inversion of the original score (1 → -1), the other relations propagated the sentiment score of the source node. By applying the relations and propagating or inverting the sentiment score accordingly, we obtained a Silver Standard of 1,293 entries. After joining the Gold and Silver Standards, we built a resource that we called *LatinAffectus*. The analysis of the outcome of this extension process reveals that propagating and reversing relations has a different impact on adjectives and nouns (see Figure 3). The former were mainly generated by reversing relations, thus coming from the list of antonyms and from the word formation rule involving the *in (neg)-* prefix, whereas the latter were mostly derived from the list of written representations, synonyms and through morphological derivations involving suffixes. The 5 most useful and productive linguistic relations we adopted were: written representations (producing 529 new lemmas), synonyms (234), *-tas/tat* (134, e.g. *concorditas* 'concord' → *concors* 'concordant'), *-ari* (124, e.g. *ordo* 'order' → *ordinarius* 'regular'), *-al* (98, *amicus* 'friend' → *amicalis* 'friendly').

6. Lexicon-Based Sentiment Analysis of a Latin Tragedy

We carried out an experiment that makes use of our resources by applying the sentiment lexicons we built to the tragedy "Medea" by Seneca, which is about Medea's revenge against the betrayer husband Jason, leading her to kill her own children.

We merged the lexicon induced with the k-NN algorithms (that is, the best performing one on the basis of our evaluation) with the Gold Standard and the Silver Standard and then we used a simple script[12] to calculate the sentiment orientation of a piece of text by summing up the scores of its words. As input, we employed the lemmatized and PoS-tagged version of the play included in the "Opera Latina" corpus.

We first calculated the polarity at the line level: out of the total number of lines of the play, 32% resulted as negative (e.g. *incognitum istud facinus ac dirum nefas* 'that unheard-of deed, that abomination', line 931) and 17% positive (e.g. *avoque clarum Sole deduxi genus* 'I shone in my noble father's light', line 210)[13]. Figure 4 shows the distribution of polar lines per character.

As a second step, we measured the polarity score of each cluster of lines (that is continuous groups of verses) pronounced by each character across the whole play: such clusters vary in length, from one line to more than 100.

Figure 5 presents the sentiment analysis throughout the tragedy, from the first to the last cluster. The predominance of the negative sentiment is evident as shown also in Figure 4. Two peaks stand out particularly, one corresponding to the first cluster, the second to the cluster number 117.

The play begins with Medea who invokes the gods of the underworlds (*noctis aeternae chaos, aversa superis regna manesque impios dominumque regni tristis* 'chaos of eternal night, realms faced away from life above, unholy spirits of the dead, lord of the gloomy realm', lines 9-11) and curses the new wife of Jason and his family (*coniugi letum novae letumque socero et regiae stirpi date* 'bring death on this new wife, death on the father-in-law and the whole royal stock', lines 17-18, *exul pavens invisus incerti laris* 'exile in fear hated and homeless', line 21).

The other peak is registered when Medea, moved by anger (*Quo te igitur, ira, mittis, aut quae perfido intendis hosti tela?* 'So where are you driving, my anger, what weapons are you aiming at your faithless enemy?, lines 916-917), decides to make her revenge (*vindicta levis est* 'the vengeance is trivial', line 901) so to punish those who betrayed her (*quaere poenarum genus haud usitatum* 'search out some exceptional kind of punishment', lines 898-899). The only evident positive peak is given by the first chorus, confirming the analysis proposed by literary critics, according to which the chorus has an antiphrastic and antithetical function compared to the prologue (Fyfe, 1983; Hine, 1989). Here gods are invoked to bless the spouses (*Ad regum thalamos numine prospero qui caelum superi quique regunt fretum adsint* 'At this royal wedding in divine support may the gods who rule heaven on high and rule the sea be present', lines 56-58) and positive adjectives are predominant (e.g. *generosus* 'noble', *mitis* 'gentle', *nitidus* 'bright').

We repeated the above described experiment, measuring the polarity score of each cluster of lines by using the lexicon developed by Chen and Skiena (2014), which is made of 936 positive tokens and 1,097 negative tokens. We assigned

[12] We modified a script originally developed for the analysis of tweets: `https://github.com/stepthom/lexicon-sentiment-analysis`.

[13] Translations taken from the edition by Frank Justus Miller (2002).

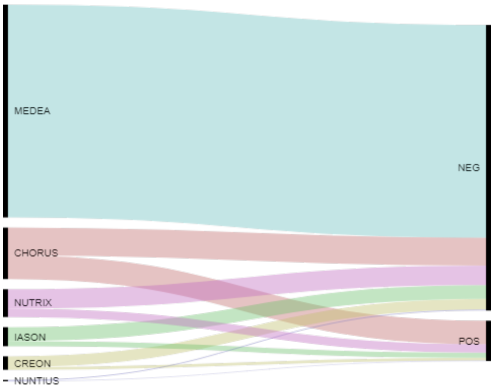


Figure 4: Distribution of negative and positive lines in the “Medea” of Seneca.

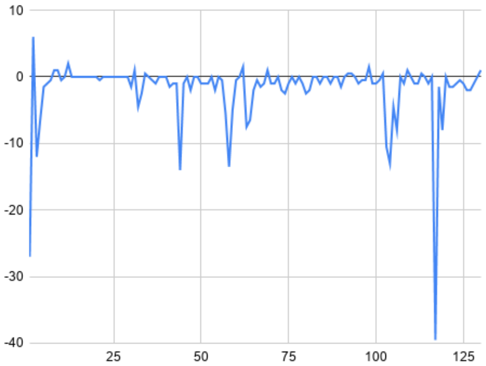


Figure 5: Sentiment orientation throughout the “Medea” of Seneca.

	PoS		SENTIMENT			
LEXICON	**ADJ**	**NOUN**	**POS**	**NEG**	**NEUT**	**TOT**
Cross-lang. Projection	3,654 (34.7%)	6,863 (65,3%)	5,259 (50.0%)	5,005 (47.6%)	253 (2.4%)	10,516
Induction (k-NN)	431 (41.8%)	599 (58.2%)	463 (44.9%)	567 (55.1%)	-	1,030
Gold Standard	454 (39.7%)	690 (60.3%)	231 (20.2%)	301 (26.3%)	612 (53.5%)	1,144
Silver Standard	512 (39.6%)	781 (60.4%)	271 (21.0%)	333 (25.7%)	689 (53.3%)	1,293

Table 6: Composition of the resources presented in this paper.

the score +1 to all the positive tokens and -1 to all the negative ones. This lexicon covers 4.5% of all the tokens in the tragedy: it is important to note that our lexicon (silver standard + induced lexicon), made of 3,253 lemmas, covers 23.7% of all the tokens and 50.8% of all nouns and adjectives. As shown in Figure 6, we obtained a majority of neutral clusters (43.1%), a greater number than we found using our lexicon (33.8%). For example, the first cluster of lines pronounced by Jason (cluster 63, lines 431-446) results as neutral whereas, with our approach, we obtained a very negative score (-7.5): it is a monologue on the cruelty of fate and on the difficult situation of Jason, featuring lemmas such as *asper* ‘adverse’, *durus* ‘hard’ and *periculum* ‘risk’. Moreover, the first chorus results as negative instead of positive; on the contrary, the most evident positive peak (+4) corresponds to the second chorus (cluster 57, lines 301-379). However, the negative score (-5.5) we obtained for the second chorus by using our lexicon seems more in line with the content of the cluster, where the chorus curses the navigation and the audacity of the Argonauts (*Audax nimium qui freta primus rate tam fragili perfida rupit* ‘Daring, too daring, the man who first split the treacherous seas with a boat so fragile’, lines 301-302), which broke the natural order of things. In addition, Medea is defined as an evil worse than the sea (*maiusque mari Medea malum*, line 362).

7. Conclusion and Future Work

In this paper we presented a new set of sentiment lexicons of Latin adjectives and nouns whose content is summarized in Table 6 and freely available online: `https://github.com/CIRCSE/Latin_Sentiment_Lexicons`. The first lexicon was generated using a cross-language projection method, the second was induced from distributed vector representations of words with a k-NN algorithm implementation. We also released a Gold Standard created by two experts of Latin language and culture, following a multi-stage process and an extensive reconciliation phase. A Silver Standard was then built by deriving new entries through synonym, antonym and derivational relations with the entries in the Gold Standard. Graphical variants of lemmas present in the Gold Standard were added to the Silver Standard as well.

Lexicons automatically assembled with the cross-language projection and the induction methods were evaluated against the Gold Standard: the k-NN algorithm applied to Latin word embeddings proved to be a promising approach for the automatic creation of sentiment lexicons without the need of any manually annotated resource. Other induction algorithms were tested as well, but achieved lower performances.

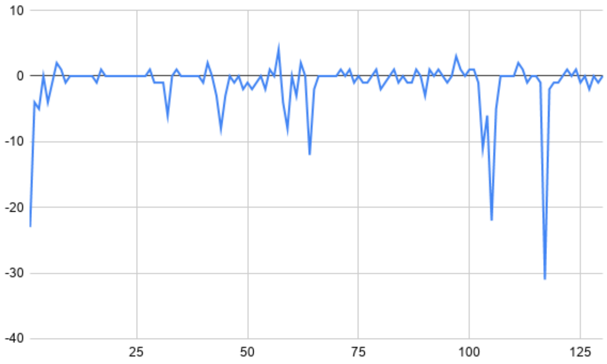


Figure 6: Sentiment orientation throughout the “Medea” of Seneca using the Chen and Skiena (2014) lexicon.

Several future works are envisaged. One of our short-term goal is to link the sentiment score of the lemmas provided by *LatinAffectus* to the corresponding lemma included in the LiLa knowledge base relying on existing ontologies designed for the semantically interoperable representation of linguistic resources for sentiment analysis (Buitelaar et al., 2013; Declerck, 2016). In addition, we plan to apply induction methods to generate time-specific sentiment lexicons and, in this way, support the diachronic analysis of Latin. Indeed, by using word embeddings trained on corpora of texts belonging to different periods, it will be possible to obtain lists of lemma-sentiment pairs specific for each period represented in the source data. A preliminary experiment run applying the k-NN algorithm to word embeddings pre-trained on 904.400 lemmas of the "Computational Historical Semantics" corpus, a manually curated collection of Latin documentary texts written between the 2nd and the 15th century AD (Jussen and Rohmann, 2015), generated a lexicon in which 71% of the entries were different from the ones obtained on the "Opera Latina" corpus. In particular, terms related to legal (e.g. *criminosus* 'guilty', *poenalis* 'penal') and Christian/Ecclesiastical issues (e.g. *abbatissa* 'abbess', *peccatrix* 'female sinner', *saluamentum* 'salvation') emerge, together with lemmas derived from German of Anglo-Saxon words, such as *faida* 'protector' and *mundiburdus* 'hostility'. Another future work is related to the ongoing extension and cleaning of Latin WordNet. Once the new version of it will be available, we will test dictionary-based induction methods on it so to obtain a sense-based sentiment resource.

8. Acknowledgments

This work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme via the "LiLa: Linking Latin" project - Grant Agreement No. 769994. The authors also wish to thank Giovanni Moretti for his technical assistance.

9. Bibliographical References

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Bamman, D., Passarotti, M. C., Busa, R., and Crane, G. (2008). The annotation guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank. The treatment of some specific syntactic constructions in Latin. In *In Proceedings of LREC 2008*, pages 71–76. ELDA.

Buitelaar, P., Arcan, M., Iglesias Fernandez, C. A., Sánchez Rada, J. F., and Strapparava, C. (2013). Linguistic linked data for sentiment analysis. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics*, pages 1–8.

Busa, R. (1974). *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ*. Frommann - Holzboog.

Castiglioni, L. and Mariotti, S. (1996). *Vocabolario della lingua latina: IL: latino-italiano, italiano-latino*. Loescher.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389.

Cho, H., Kim, S., Lee, J., and Lee, J.-S. (2014). Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. *Knowledge-Based Systems*, 71:61–71.

Declerck, T. (2016). Representation of polarity information of elements of german compound words. In *LDL 2016 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources*, page 46.

Denooz, J. (2004). Opera Latina: une base de données sur internet. *Euphrosyne*, 32:79–88.

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5.

Fillmore, C. J. e. a. (1982). Linguistics in the morning calm. *Linguistics Society of Korea. Frame Semantics. Seou: Hanshin*.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V., and Zampedri, F. (2019). Nunc Est Aestimandum. Towards an Evaluation of the Latin WordNet. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2029)*. CEUR-WS. org.

Fyfe, H. (1983). An analysis of Seneca's Medea. *Ramus*, 12(1-2):77–93.

Hamilton, W. L., Clark, K., Leskovec, J., and Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.*, volume 2016, page 595. NIH Public Access.

Hine, H. (1989). Medea versus the Chorus: Seneca "Medea" 1-115. *Mnemosyne*, 42(Fasc. 3/4):413–419.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Jussen, B. and Rohmann, G. (2015). Historical Semantics in Medieval Studies: New Means and Approaches. *Contributions to the History of Concepts*, 10(2):1–6.

Litta, E., Passarotti, M., and Culy, C. (2016). Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the third Italian conference on computational linguistics (clic–it 2016)*, pages 185–189.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

Mihalcea, R., Banea, C., and Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of* *the association of computational linguistics*, pages 976–983.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184.

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Stoyanov, V., and Zhu, X. (2016). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation*, 50(1):35–65.

Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2009). Sentiful: Generating a reliable lexicon for sentiment analysis. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Passarotti, M., Budassi, M., Litta, E., and Ruffolo, P. (2017). The Lemlat 3.0 Package for Morphological Analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, number 133, pages 24–31. Linköping University Electronic Press.

Passarotti, M. C., Cecchini, F. M., Franzini, G., Litta, E., Mambrini, F., and Ruffolo, P. (2019). The LiLa Knowledge Base of Linguistic Resources and NLP Tools for Latin. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, pages 6–11. CEUR-WS. org.

Seneca. (2002). *Tragedies, Volume I: Hercules. Trojan Women. Phoenician Women. Medea. Phaedra*. Cambridge, MA: Harvard University Press, Loeb Classical Library 62 edition.

Sidarenka, U. and Stede, M. (2016). Generating Sentiment Lexicons for German Twitter. In *PEOPLES 2016*, page 80.

Sidarenka, U. (2019). *Sentiment analysis of German Twitter*. doctoralthesis, Universität Potsdam.

Simpson, D. P. (1959). *Cassell's Latin dictionary*. Simon Schuster Macmillan Company.

Skřivan, A. (1890). *Latinská synonymika pro školu i dum*. V CHRUDIMI.

Souter, A. (1968). *Oxford Latin dictionary: OLD*. Clarendon Press.

Sprugnoli, R., Passarotti, M., and Moretti, G. (2019). Vir is to Moderatus as Mulier is to Intemperans. Lemma Embeddings for Latin. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: studies using the General Inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241–256. ACM.

Takamura, H., Inui, T., and Okumura, M. (2005). Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 133–140. Association for Computational Linguistics.

Velikovich, L., Blair-Goldensohn, S., Hannan, K., and McDonald, R. (2010). The viability of web-derived polarity lexicons. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 777–785. Association for Computational Linguistics.