Are Proverbs the New Pythian Oracles? Exploring Sentiment in Greek Sayings

Anonymous ACL submission

Abstract

Proverbs are among the most fascinating linguistic phenomena that transcend cultural and linguistic boundaries. Yet, much of the global landscape of proverbs remains underexplored, as many cultures preserve their traditional wisdom within their own communities due to the oral tradition of the phenomenon. Taking advantage of the current advances in Natural Language Processing (NLP), we focus on Greek proverbs, analyzing their sentiment. Departing from an annotated dataset of Greek proverbs, we expand it to include local dialects (ντοπιολαλιές), effectively mapping the annotated sentiment. We present (1) a way to exploit LLMs in order to perform sentiment classification of proverbs, (2) a map of Greece that provides an overview of the distribution of sentiment, (3) an combinatory analysis in terms of the geographic position, dialect, and topic of proverbs. Our findings show that LLMs can provide us with an accurate enough picture of the sentiment of proverbs, especially when approached as a non-conventional sentiment polarity task. Moreover, in most areas of Greece negative sentiment is more prevalent.

1 Introduction

Proverbs are a unique linguistic phenomenon that carries cultural significance, metaphorical wisdom, and timeless relevance. The latter characteristic can be seen as the symptom of the subjectivity of interpretation. Marti Solano and Rondinelli (2021) discuss how relevance is timeless while meaning can change, with one of their use cases being *A rolling stone gathers no moss*. While the proverb once urged people to remain in one place and build a steady future, with *moss* symbolizing stability, the metaphoric *moss* has more recently taken on a negative meaning, suggesting something that holds you back from exploring new opportunities. Studies like this are parts of two 'sibling subfields' that are part of broader fields like linguistics, philology, and folkloristics, called *paremiology* and *paremiography*, and which are concerned with the study, collection and classification of proverbs, as well as addressing questions on their definition, form, structure, style, content, function, meaning and value (Mieder, 2004).

The combination of paremiology (and/or paremiography) and Natural Language Processing (NLP) is still quite sparse. Most paremiology studies focus on a handful of use cases of proverbs with a very narrowed-down focus such as pragmatical or social aspects, as in the case of Lomotey and Csajbok-Twerefou (2021), while NLP studies in most cases utilize proverbs for their metaphorical aspect to help in broader tasks, such as metaphor detection or translation (Özbal et al., 2016; Romaniuk-Cholewska, 2024; Wang et al., 2025). The limited overlap is also driven by a major obstacle: proverbs, being rooted in oral tradition, are difficult to date, making it challenging to determine when they first got established (Khanaifsawy, 2023). This, in turn, complicates the task of compiling data for large-scale paremiological analysis. Another inhibiting factor is the 'tautological' tendency of technology—developing technology for its own sake—often at the expense of applying it to the digital humanities or to more traditional linguistic research.

Recognizing the importance of cultural heritage and the opportunities we are gifted with due to the current NLP advances, we attempt to draw the paremiological landscape of Greece from a sentiment perspective. As noted by Gwóźdź (2016), "proverbs are emotionally saturated … and a close examination of the valency of paremic context may provide an insight into the usage of particular adages". At the time that chapter was written, however, the author acknowledged that the state of AI was not advanced enough to achieve satisfactory accuracy. In this work, we revisit the task in light of the progress brought by the LLMs. More-

over, we highlight that the proverbial landscape of Greece (similar to that of many medium- and low-resourced languages) remains largely understudied, particularly with regard to its local varieties (ντοπιολαλιές). To address this gap, we provide contrastive maps that not only serve as research bases for the study of the sentiment of proverbs but also establish a foundation for further research.

Although one of the most definitive aspects of proverbs is their structural rigidity, paradoxically enough, they also appear in multiple variants, a characteristic that makes proverbs tough to detect and a challenge for corpus studies (Norrick, 2015; Kljajevic and Šarić, 2025). Moreover, the need to preserve an endangered language or dialect is a strong enough reason to elevate the value of proverbs in any language (Lauhakangas, 2015).

Considering the aforementioned challenges, our study is guided by one main research question (RQ_m) and three supporting sub-questions (RQ_{sub1-3}):

RQ_m: How do sentiment and topic vary across regions in the Greek proverbial landscape?

- RQ_{sub1}: Can we aggregate and map the sentiment of proverbs to specific geographical locations?
- RQ_{sub2}: What are the primary challenges in accurately classifying sentiment in proverbs?
- RQ_{sub3}: How accurately can LLMs help us with such an analysis?

Our findings operate on three levels. First, in terms of annotation and subjectivity, sentiment and emotion cannot be treated as objective indicators, so caution is needed when using proverbs for tasks like metaphor detection. Second, in mapping local varieties, negative sentiment dominates, with topics often linking agricultural life to hardship, pain, and forces beyond human control. This contrasts with urban centers such as Athens and Thessaloniki, where data is notably almost non-existent. Third, LLM-based methods can be highly valuable for paremiology research, since they can achieve up to 70% F_1 scores and show strong correlation with human annotations

The rest of the paper is structured as follows. We begin by introducing related work (§2), covering both traditional paremiology and NLP approaches to sentiment, thematic analysis and categorization, as well as other important studies on proverbs. We then present our methodology (§3), which leads to the two main result sections: agreement analysis (§4.1) and model experiments (§4.2). These are followed by a discussion (§5) in which we highlight key findings and their implications for paremiological analysis. Finally, we conclude with some remarks and outline directions for future work (§6), limitations and ethical implications.

2 Related Work

Proverbs have been a tantalizing research target for many different disciplines. Folklorists see proverbs as a major carrier of folk wisdom and traditional lore; linguists focus on the different linguistic components, from syntax to parallelisms and imagery; corpus linguists focus on the context of proverbs, while lexicographers focus on their registration and categorization (Norrick, 2015). All these perspectives can, of course, interact and provide very useful insights into intra- and cross-cultural comparisons. This interaction is evident in this paper as well, as we use computational methods to discover ethnographic and linguistic patterns through the use of proverbs. This section is mainly informed from literature found in NLP and linguistics.

2.1 Proverbial Sentiment

Proverbs and sentiment analysis are often seen as complementary tasks, but with the main target being the improvement of sentiment analysis. One exception is Ghosh and Srivastava (2022) whose particular focus is on creating a corpus of narratives and contexts for proverbs, and where they perform a small scale sentiment analysis to ensure the diversity of the narrative. Comparatively, more work has been conducted on idioms and sentiment rather than proverbs and sentiment, as in the case of Williams et al. (2015). The authors show that the inclusion of idioms as features may potentially improve the results of traditional sentiment analysis. Idioms and sentiment are also explored in Jochim et al. (2018) and Tahayna et al. (2022). The first build a sentiment lexicon called the Sentiment Lexicon of Idiomatic Expressions (SLIDE). The second propose a pipeline for automatic annotation of sentiment in idiomatic phrases by using data augmentation proceeding to a full replacement of an idiom with its factual meaning or definition retrieved from an external knowledge source (and expanded version of SLIDE) and fine-tuning BERT to perform the classification. Hwang and Hidey (2019) adopt a

different approach where they examine the compositionality of idioms using SLIDE with sentiment scores from WordNet and the Dictionary of Affect in Language, finding no clear relation between component- and phrase-level sentiment, supporting the view that idioms are non-compositional in sentiment as well as semantics.

Despite all the above efforts, the sentiment of proverbs for cultural, linguistic, and ethnographic reasons remains much unexplored. Not only do we see that in most cases idioms are examined more frequently that proverbs (although one can include the other), but in most cases the research is in English, and focuses on the improvement of sentiment analysis rather than the linguistic analysis of the expressions. This paper attempts to fill these gaps.

2.2 Proverbial Thematics and Categorization

The thematology and categorization of proverbs can still face a level of arbitrariness and are usually organized according to the compiler's interests or interpretative approach. These classifications can serve different purposes, such as highlighting national traditions or building scientific systems of organization. For example, Vladimir Dalj's Russian collection is a thematically arranged work that clearly reflects a national aim (Dalj, 2003).

In scientific categorization, the guiding principle is that categories should not overlap (Lauhakangas, 2015). Yet, proverbs are rarely one-dimensional, and it is often difficult to assign a single category to every expression. The most ambitious and internationally recognized framework, however, is the Matti Kuusi International Type System of Proverbs, developed by Matti Kuusi and further elaborated by Outi Lauhakangas (Lauhakangas, 2001, 2013). This system is based on three elements: the extensive collection of proverb materials preserved in the folklore archives of the Finnish Literature Society in Helsinki, a digital database of international proverb types with literature references, and a comprehensive thematic and structural classification. The thematic range of the system is broad, covering areas such as practical knowledge of nature, faith and basic attitudes, fundamental observations about human life and society, morality, social life and interaction, communication, norms and agreements, learning and coping, and even perceptions of time. However, it comes with limitations, i.e., treating language groups as broad categories, making it easier to compare across cultures but meaning that local dialects could be glossed over.

2.3 What about Greek Proverbs?

NLP studies on Greek proverbs are quite rare. Pavlopoulos et al. (2024) introduce a publicly-available dataset with more than one hundred thousand Greek proverb variants. They also proceed to attribute classification, showing that specific locations are classified with higher accuracy than others, while also attributing more than 3k unregistered Greek proverbs using conformal prediction to estimate their most likely origins with mathematically guaranteed confidence. At the low-resource language frontier, Dimakis et al. (2025) use proverbs for dialect normalization using LLMs, while showing that previous experiments on those proverbs ignore deeper dialectal attributes.

3 Method

Our methodology consists of three phases. The first phase concerns the collection and annotation of proverbs for sentiment analysis. In this phase, we also analyze the inter-annotator agreement and the distribution of labels as a form of quality assurance for our dataset development. The second phase employs prompting techniques with LLMs to predict sentiment in our new dataset and to evaluate different types of prompting settings and models. The final phase involves an upscaled inference setting for the geographical visualization of the proverbial map of Greece in terms of sentiment. The local variability of proverbs makes their registration and analysis a challenging task (Norrick, 2015), which we believe can benefit greatly from recent advances in LLMs.

3.1 Dataset Development

Data Collection. In this research, we use two types of proverbs. For the first type, we use the term *standard*, which corresponds to modern Greek without any recognised dialectal attributes. These proverbs were collected from *gnomikologikon*, an online collection of proverbs, sayings, and quotes.[1] The second type are proverbs with dialectal attributes or 'locality', therefore we refer to these proverbs as *localized* proverbs. An example of the two types can be found in Table 1. The localized proverbs were the ones compiled by Pavlopoulos et al. (2024), which come from the Hellenic Folklore Research Centre.[2][3]

[1] https://www.gnomikologikon.gr/
[2] https://kentrolaografias.gr/
[3] The data are under a CC BY-NC-ND 4.0 licence.

Standard	Localized (Naxos Dialect)
της νύχτας τη δουλειά τη βλέπει η μέρα και γελά	τση νύχτας τσοί δουλειές τσοί βλέπεις μέρα και γελάς

Table 1: Example comparing a Standard Greek proverb with a Naxian dialect proverb. The word-by-word translation is *the day sees the work of night and laughs*, meaning that whatever is done at night is done without much attention.

Annotation Procedure. We begin by randomly selecting a set of 200 standard proverbs as a pilot sample to generate an initial list of potential emotions that the annotators could insert as free-text. We later map the identified emotions to sentiment to further verify the accuracy and consistency of our sentiment annotations, as described in §4.1. As discussed in §2, there is currently no established framework for classifying the emotions expressed in proverbs. Existing taxonomies, such as Maati Kuusi's proverb type system (Lauhakangas, 2013), treat many emotions (e.g., pride, humility) primarily as thematic categories rather than as affective labels. The pilot annotation phase yielded 47 emotions, which can be found in Appendix A, along with the guidelines. The pilot proverbs are not used again for the purposes of this study.

For the next step, we select 300 new unseen instances of standard proverbs (as we did not want to bias the annotators that participated in the pilot experiment) for a new round of annotation that required the annotators to label emotion and sentiment per proverb. The guidelines can be found in Appendix A. We did not opt for a direct annotation of the proverbs with locality, because they could be hard to understand if someone is not familiar with the dialect, and which could lead to loss of meaning or inaccurate results. Moreover, finding native speakers of some dialects can be hard, especially in the case of proverbs that their origin location does not officially exist anymore (e.g., Anatolia). The results of the analysis of the agreement on both of these aspects is discussed in §4.1.

Annotator Profiles. We recruit 13 annotators that are postgraduate students and native Greek speakers, possessing advanced knowledge in language technology and with prior annotation experience. The gender distribution was fairly balanced, and all participants took part in the annotation project voluntarily.

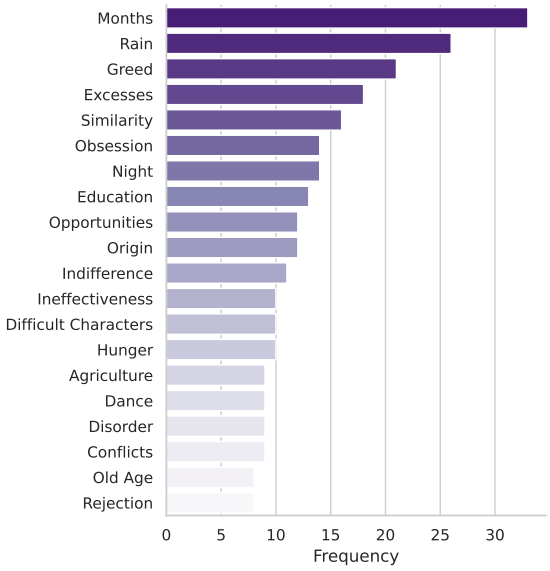

Figure 1: Top 20 topics according to the labels found in the *gnomikologikon*. The topics are translated from Greek to English.

3.2 Topic Statistics

The standard proverbs we collected were pre-categorized by the online source, therefore we leveraged these topic assignments in our research. In Figure 1, we see the top 20 most frequent topics. With the months being the most common topic, we are reminded that proverbs are verbalized wisdom that most often refers to common hardships, and which consequently is rich with agricultural themes, such as the seasons and time, which were pivotal to the survival of the people. Another observation, which supports our claim that no separate taxonomy of emotions exists for proverbs, is that many topics we encounter here, such as greed, obsession, and indifference, can also be interpreted as emotions and are closely related to those identified in the pilot annotation, such as desire, disappointment, and malice (see Appendix A).

3.3 Assigning Silver Labels

Using the 300 standard Greek proverbs, we evaluate different prompting techniques in order to get the best possible sentiment prediction. We then map the standard proverbs to the localized proverbs (11.5k instances), using Levenshtein edit distance. An example of visualizing the differences and simulating edit distance can be found in Appendix D. We chose Levenshtein distance for our similarity metric as dialect variations usually involve minor phonetic or orthographic changes and although it

might be written differently in spelling, the words remain similar. Since edit distance measures literal character changes (insertions, deletions, substitutions), it is sensitive to small changes while preserving word order, as rigidity is one of the most characteristic attributes of proverbs. This yielded 430 gold-standard localized proverbs with labels for sentiment as well as the coordinates of where the proverb is registered, which were included in the original dataset of Pavlopoulos et al. (2024).[4]

3.4 In-Context Learning

We performed few-shot learning, using both open-source and close-source models, including two specifically designed for Greek language tasks. We used different prompting techniques with different models for the evaluation. In an attempt to have a variety with regard to models, we chose relatively small LLMs (Krikri-8B, Meltemi-7B, Mistral-7B) as well as two larger ones (GPT-4o and Llama-70B). The full prompts are provided in the Appendix. We included variations of zero-shot settings, summarized as follows:

- **Simple Zero-shot (Z_0):** For an input instance x, the model is prompted directly to predict the label $\hat{y}$ without any example provided.
- **Batch Prompting (Z_b):** A set of n instances $\{x_1, x_2, \ldots, x_n\}$ is given in a single prompt, where $n \in \{10, 20, 30\}$. The model predicts the corresponding labels $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$.
- **Percentage Prediction Prompting (Z_p):** Instead of directly predicting a label, the model is asked to output the probability distribution $\hat{p}(y \mid x)$. This provides a measure of uncertainty and allows for further error margin in evaluation.

We also experiment with few-shot setups, where each class is provided with $k \in \{1, 2, 3\}$ examples (shots) in combination with the following shot selection strategies:

- **Random Pick (RP):** Select n instances randomly.
- **Low-Variance Selection (LVS):** Choose n examples minimizing label variance, that is where annotators agree the most:

$$S^* = \arg\min_{|S|=n} \mathrm{Var}\{y_i : i \in S\}.$$

[4]Some of the coordinates were missing, especially in the case of areas that are not officially present. In that case, we traced the area and the coordinates manually.

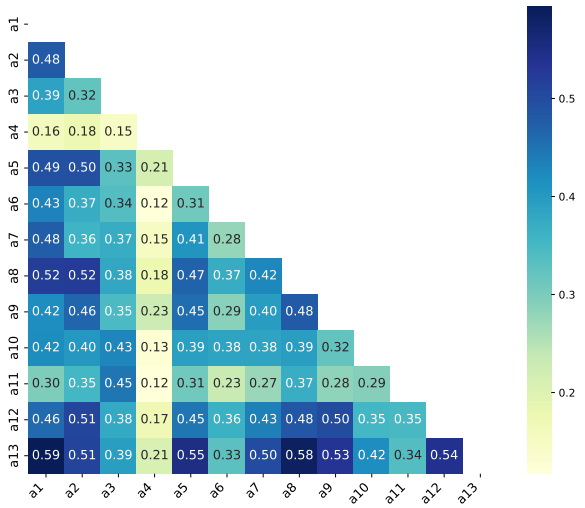

Figure 2: Pearson correlation heatmap in the sentiment annotations. From a1 to a13 are the annotators.

Metric	Sentiment	Emotion	EmoSent
κ	25.2%	08.5%	24.0%
α	24.9%	08.1%	23.6%
Spearman	39.2%	07.7%	36.1%
Pearson	40.8%	06.9%	37.0%

Table 2: Agreement metrics. From top to bottom: Cohen's κ, Krippendorf's α, Spearman and Pearson correlations for sentiment, emotion, and emotion mapped to sentiment (EmoSent column)

- **Dynamic (DYN):** Choose the top-n examples from the LVS list that are closest, based on an embedding space, to the proverb in question (i.e., using cosine similarity to fetch k shots).

4 Results

4.1 Agreement Analysis

Both with regard to sentiment and emotion, the annotation presents low IAA, possibly as a result of the highly metaphorical language and interpretation. The summary of the IAA metrics for sentiment and emotion is shown in Table 2. One of the annotators had to be removed due to high disagreement and small correlation with the other annotators, as illustrated in Figure 2. More information on the correlations can be found in Appendix B. The emotion-based IAA is much lower, possibly due to the 47 labels the annotators were provided with, which can render the task overwhelming. One important observation, however, is that when we map the emotion labels to the sentiment labels (dubbed as EmoSent), IAA is very close to the one of sentiment where the annotators had to choose only

Greek Proverb	English Translation
Η χάρη θέλει αντίχαρη και πάλι χάρη είναι.	A favor wants a return favor, and yet it is still a favor.

Table 3: Example of an ambiguous case or a proverbial paradox.

out from 3 labels. The robustness of the sentiment mapping is further supported by the analysis of overlapping examples between the two datasets. Specifically, we measured self-agreement by taking the overlapping examples that each annotator had labeled in both annotation tasks (sentiment and emotion), mapping their emotion labels to sentiment categories, and then checking whether their sentiment decision was consistent across the two instances. For each annotator, the proportion of consistent labels was computed as a percentage, and these percentages were then averaged to obtain an overall self-agreement score, which yielded a score of 72.33%.

Transforming into an Ordinal Problem Since the agreement analysis so far has yielded a fairly low IAA, and given that more and more NLP approaches adopt a more prescriptive paradigm when annotating and examining data (Fleisig et al., 2024; Romberg et al., 2025), we transform the task into an ordinal regression problem only. Specifically, we process the sentiment annotations from the 13 annotators per proverb and we compute the number of negative, ambiguous, and positive votes per instance. These counts are then normalized to obtain sentiment distributions, from which a soft sentiment score is calculated by taking the difference between positive and negative votes divided by the number of the annotators. If at least 10 people agree on negative or positive or ambiguous, the sentiment is labeled 'strongly' in that direction. If 7 or more agree while still being the dominant sentiment, it's labeled 'mostly'. The label distribution for the new labels can found in the Appendix B.

Reminiscent of an oracle, we come to the conclusion that maybe it is not of our benefit talking about the sentiment of the proverbs as a legitimate (classification) task. This can be shown by examining the ambiguous cases that appear during this analysis. We present an example in Table 3. The example showcases a proverbial case of a paradox, where resolution does not work (Norrick, 2015). The most prevalent meaning of the proverb is that when someone does you a favor, you ought to return it but even when returned, it is still a favor, not a debt. The paradox in that is that it every favor expects repayment, but even that repayment counts as another favor. Therfore, we come with two interpretations, one altruistic (positive), namely a selfless act of kindness can lead to an endless chain of favors and one cynical (negative), namely you are expected to return the favor.

4.2 Evaluating LLM Techniques

We use the 300 annotated standard proverbs to examine which prompting technique is best when predicting sentiment on proverbs. All the results are presented in Tables 4, 5, and 6.

Zooming out, we see that, as expected, the larger models perform better. GPT-4o outperforms the others in almost all settings: in the zero- and few-shot setups it reaches up to 0.69 macro-F_1, shows very small errors and strong correlation in the percentage prediction setting, and also achieves the best performance in batch prompting with an F_1 score of up to 0.63.

However, we also see that Krikri and Llama keep up quite well. Krikri, despite its relatively few parameters—and likely thanks to its Greek fine-tuning—is the runner-up in the zero- and few-shot settings, while Llama performs almost on par with GPT-4o in the percentage prediction task.

In terms of the best setting, given the subjectivity of the task, we decide to approximate our best open-source model (Llama-70b) using the Z_0 setting for our map. In this way, we have a greater error margin without losing credibility and we can paint the map of sentiment of Greek proverbs respecting both sentiments that might be reflected in our proverbial instances.

4.3 Predicting on Localized Proverbs

In order to obtain the sentiment maps of the local variety proverbs, we first predict on the 430 instances that we mapped, as described in Section 3. These ground-truth maps can be found in the Appendix E. However, due to the limited number of regions covered (105), we proceed to predict on the whole dataset of 11.5k proverbs. The results can be found in Figure 3. Although the maps are scaled up by predicting from 300 to 11.5k proverbs, the scope of regional analysis is still restricted: the total number of regions increases only to 285, and these correspond to smaller parts of larger areas rather than independent county-like administrative divisions, highlighting the need for further research.

Model	0-shot	1-shot			2-shot			3-shot		
		RP	LVS	DYN	RP	LVS	DYN	RP	LVS	DYN
KriKri-8B	0.17	0.64	0.63	0.61	0.56	0.51	0.60	0.44	0.52	0.62
Meltemi-7B	0.04	0.44	0.44	0.60	0.36	0.42	0.57	0.30	0.36	0.48
Mistral-7B	0.38	0.57	0.49	0.33	0.54	0.59	0.30	0.52	0.53	0.33
GPT-4o	0.63	0.64	0.64	0.66	0.66	0.65	0.68	0.63	0.68	0.69
Llama3-70B	0.47	0.55	0.57	0.53	0.62	0.57	0.56	0.59	0.59	0.58

Table 4: Weighted-F_1 scores for each model across different shot settings.

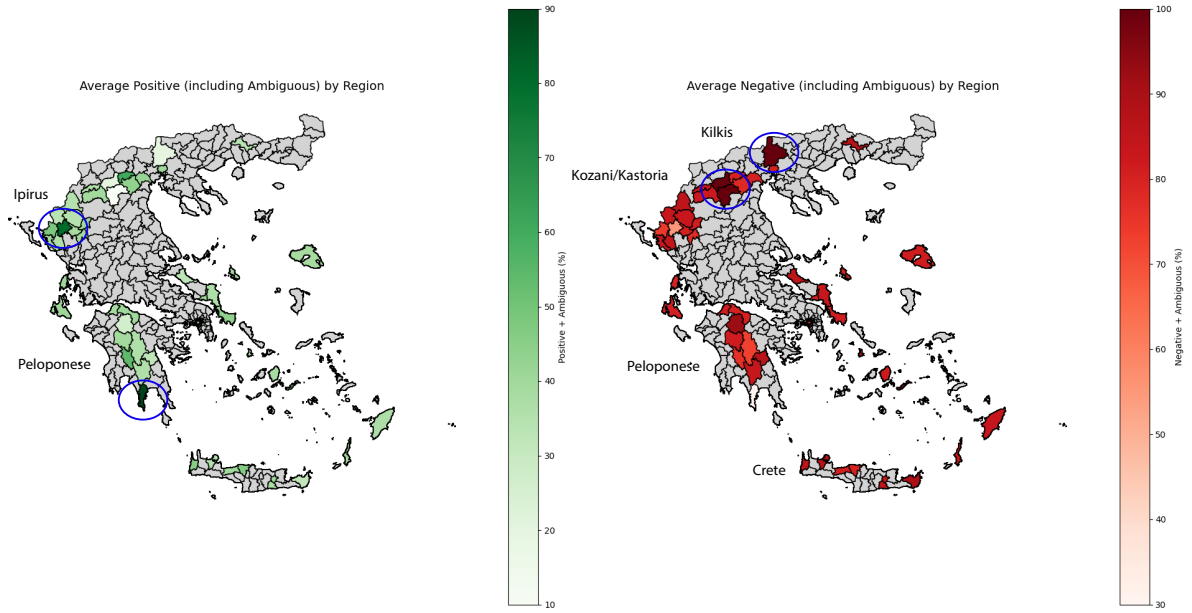

Figure 3: Predicted positive and negative sentiment maps of Greece. In both cases, ambiguous instances are aggregated into the positive and negative classes. Interactive versions of the maps, including proverb counts and the option to view individual proverbs, will be released upon acceptance.

Model	MAE	MSE	ρPos	ρAmb	ρNeg
KriKri-8B	0.30	0.15	0.58	0.28	0.54
Meltemi-7B	0.34	0.20	0.05	-0.01	0.05
Mistral-7B	0.30	0.15	0.25	-0.04	0.19
GPT-4o	0.17	0.05	0.71	0.42	0.71
Llama3-70B	0.19	0.07	0.71	0.31	0.65

Table 5: Mean absolute error (MAE) and mean squared error (MSE) for percentage prediction prompting (Z_p), along with Pearson correlations (ρ) for each predicted class (positive, ambiguous, negative).

Model	10-batch	20-batch	30-batch
KriKri-8B	0.31	0.34	0.32
Meltemi-7B	0.27	0.27	0.24
Mistral-7B	0.34	0.38	0.40
GPT-4o	0.63	0.64	0.58
Llama3-70B	0.33	0.38	0.36

Table 6: Batch prompting weighted-F_1 scores per model

Still, significant observations can be made. Looking at both maps, it is evident that negative proverbial sentiment is more intense than positive sentiment. Both maps present some regional variation with regard to sentiment distribution. For example, we see that the area close to Ipirus (northwest), as well as the south of the Peloponese, there is more positive sentiment (circled parts). On the negative map, we observe that Crete, as well as central Peloponese and especially parts of Northern Greece (around Kozani, Kastoria, and Kilkis) are particularly negative.

Another important observation is the fact that there is a lack of proverbs around urban centers such as Athens and Thessaloniki, and not even in the vicinity of the metropolitan areas, like the Attica region. This sparsity may reflect that proverbs often emerge from rural communities where oral traditions play a central role in everyday communication (Arewa and Dundes, 1964). In contrast, urban areas tend to exhibit greater population mo-

bility, cultural mixing, and reliance on formal education, which might hinder a more metaphorical, poetically licensed use of language.

Topic Inspection A manual examination of the topics of the most positive and most negative proverbs looks in accordance with the topic distribution that we see in Section 3 in Table 1, with 'Month' and 'Rain' being the most prevalent. Though these are potentially associated with the positive sentiment, such as in the case of Αν βρέξει ο Απρίλης δυο νερά κι ο Μάης άλλο ένα, χαρά σ' εκείνον το ζευγά που 'χει πολλά σπαρμένα. translated in English as if *April rains twice and May another time, joy to the farmer who has many crops sown*. In contrast, the negative sentiment is associated with topics that correspond to emotions such as 'Rejection', 'Ingratitude', 'Excuses', but also more religious topics such as 'Devil'.

5 Discussion

On the Research Questions. The sentiment map of Greece can be characterized as distinctly non-homogeneous, as it exhibits variations of positive and negative sentiment across different regions, with negative sentiment appearing more prevalent overall. Advances in NLP now make it feasible to generate such maps, but further work is needed in terms of evaluation and validation. Moreover, these maps should not rely solely on polarity-based labeling—whether automatic or human—as such tasks can be highly subjective. Instead, they should aim to capture the full spectrum of sentiment and all possible outcomes. Therefore, evaluation remains the primary challenge, compounded by the scarcity of data for more informed analysis and the limited availability of metadata, such as location information.

On the Origin of Greek Proverbs, An analysis such as the one presented in this paper cannot give us any information about the genesis of a specific proverb, but rather its spreading. An attempt to discover the origin of any proverb is very much futile, as thoroughly described in much of the paremiological literature (Norrick, 2015). This is because it is most likely that proverbs are coined by a single individual, who managed to grasp a universal truth in an adequate yet metaphorical manner, and was expressive enough that it was adopted and used by the folk. This however and as discussed in previous sections implicated that the way the proverbs are used has led to variation. In this specific work, although we map the proverbs and their sentiment to specific regions in Greece, the only indication that we have that these proverbs belong to the specific regions are the dialectal properties of the proverbs. This however does not mean that one region 'owns' a proverb, but rather it owns the specific variety of the proverb. More research should look into the differences between the local varieties both regarding linguistic differences but also topic, emotion, and sentiment.

On the Use of NLP for Paremiology. This study showcased one method to delve further into the study of paremiology using modern technological tools. As in many NLP tasks, evaluation remains a key challenge, which can limit the application of NLP tools in digital humanities, folkloristics, ethnographic research, and even more traditional linguistic studies. In this paper, we aim to open a path for further research on proverbs and cultural heritage, exploring how computational methods can help analyze semantic patterns, sentiment, regional variations, and cross-cultural similarities.

6 Conclusions

In this paper, we used NLP methods to address the task of sentiment analysis in Greek proverbs, examining at the same time local varieties. Our method involved examining the subjectivity of proverbs despite their seemingly 'fixed' nature. For that reason, we propose examining sentiment from both the perspectives of positivity and negativity as a non-mutually exclusive task. Our results show that there are, in fact, regions that are more positive and more negative, but the striking issue is the lack of data around urban areas, as well as the common themes of rain, nature, and the seasons of the year, which reflect the agricultural attitudes of the people.

Future work will take a look at the present day. As long as language and technology are evolving and data are becoming more accessible, we get to register and curate a process of how linguistic and folkloristic elements such as proverbs are evolved. We already have filtering schemas that can distinguish proverbs from other types of folklorisitic, literary, or linguistic mechanisms, such proverbial utterances, and although the lines can be burry, with the help of new computational means we can delve into more cognitive as well as social patterns of how proverbial elements evolve or disappear.

Limitations

This study is not without limitations. The presence of the same proverbs across different localities does not mean necessarily that sentiment is perceived in the same way. On the contrary, it suggests that sentiment and emotion may vary depending on the place, the background and the experiences of the peoples. This highlights, once again, the inherently subjective nature of interpreting proverbs. Additionally, while some proverbs can be considered 'gold' and be very similar in all the different varieties (such in the Levehnstein distance setting that we used to map the proverbs), others can be similar in meaning but different in form; yet both types reflect the oracle-like nature of these sayings. Their evaluation, however, is necessarily filtered through our own subjectivity, since no objective truth can be applied in this context. Finally, it is important to note that we do not know the extent to which these proverbs are actively used today, leaving their current relevance somewhat uncertain.

Ethics Statement

The participation in the annotation was completely voluntary and anonymous. All proverbs used in this study were collected from publicly available sources and existing corpora. No private or sensitive data was used. We acknowledge that annotators' interpretations of sentiment in proverbs may reflect cultural or regional biases. This research is intended for academic purposes and cultural understanding, therefore, we acknowledge that automated sentiment predictions should be used carefully and with respect to the local (proverbial communities) and definitely not in order to stereotype one community as positive and another as negative.

References

E. Ojo Arewa and Alan Dundes. 1964. Proverbs and the ethnography of speaking folklore. *American Anthropologist*, 66(6, Part 2: The Ethnography of Communication):70–85.

Vladimir Dalj. 2003. *Poslovitsy russkogo naroda [Proverbs of Russian People]*. Moscow. Originally published in 1862.

Antonios Dimakis, John Pavlopoulos, and Antonios Anastasopoulos. 2025. Dialect normalization using large language models and morphological rules. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23696–23714, Vienna, Austria. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Maja Gwóźdź. 2016. Sentiment analysis of proverbs in american discourse: A corpus-based study. In *From Motion to Emotion: Aspects of Physical and Cultural Embodiment in Language*, pages —. —.

Alyssa Hwang and Christopher Hidey. 2019. Confirming the non-compositionality of idioms for sentiment analysis. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 125–129, Florence, Italy. Association for Computational Linguistics.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Abdullah Najim Abd Al Khanaifsawy. 2023. A pragmatic study of proverbs: Sociocultural perspective. *International Journal of Research in Social Sciences and Humanities*, 13(II). P-ISSN: 2454-4671.

Vanja Kljajevic and Ljiljana Šarić. 2025. Corpus-based investigation of proverbs: Challenges and new directions. *Corpus Pragmatics*, 9:135–163.

Outi Lauhakangas. 2001. *The Matti Kuusi International Type System of Proverbs*. Number 275 in FF Communications. Academia Scientiarum Fennica, Helsinki.

Outi Lauhakangas. 2013. The matti kuusi international database of proverbs. *Oral Tradition*, 28/2:217–222.

Outi Lauhakangas. 2015. 3 categorization of proverbs. In Hrisztalina Hrisztova-Gotthardt and Melita Aleksa Varga, editors, *Introduction to Paremiology*, pages 49–67. De Gruyter Open Poland, Warsaw, Poland.

Benedicta Adokarley Lomotey and Ildiko Csajbok-Twerefou. 2021. A pragmatic and sociolinguistic analysis of proverbs across languages and cultures. *Journal of Pragmatics*, 182:86–91.

Ramon Marti Solano and Paolo Rondinelli. 2021. 'Comparisons are odious': how comparable are equivalent proverbs across languages? *Proceedings of the 14th Interdisciplinary Colloquium on Proverbs*.

Wolfgang Mieder. 2004. *Proverbs: A Handbook*. Greenwood Folklore Handbooks. Bloomsbury Academic.

Neal R. Norrick. 2015. 1 subject area, terminology, proverb definitions, proverb features. In Hrisztalina Hrisztova-Gotthardt and Melita Aleksa Varga, editors, *Introduction to Paremiology*, pages 7–27. De Gruyter Open Poland, Warsaw, Poland.

John Pavlopoulos, Panos Louridas, and Panagiotis Filos. 2024. Towards a greek proverb atlas: Computational spatial exploration and attribution of greek proverbs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11842–11854. Association for Computational Linguistics.

Dominika Romaniuk-Cholewska. 2024. The machine translation (mt) of proverbs in the eng-pl language pair. *Studies in Logic, Grammar and Rhetoric*, 69:553–570.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025. Towards a perspectivist turn in argument quality assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

Bashar M. A. Tahayna, Ramesh Kumar Ayyasamy, and Rehan Akbar. 2022. Automatic sentiment annotation of idiomatic expressions for sentiment analysis task. *IEEE Access*, 10:122234–122242.

Minghan Wang, Viet Thanh Pham, Farhad Moghimifar, and Thuy-Trang Vu. 2025. Proverbs run in pairs: Evaluating proverb translation capability of large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1646–1662, Vienna, Austria. Association for Computational Linguistics.

Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. 2015. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. Prometheus: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3787–3793. European Language Resources Association (ELRA).

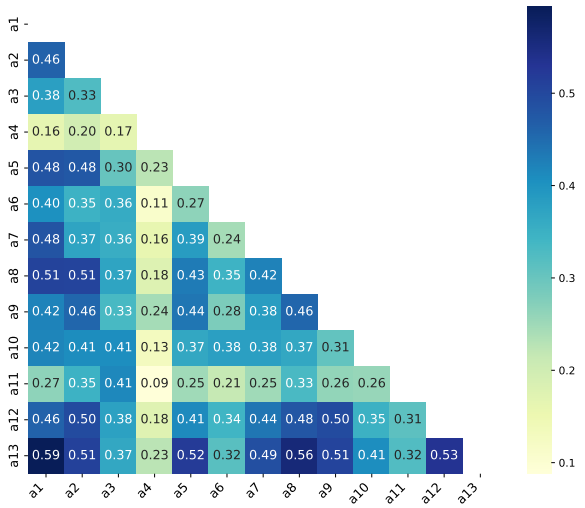

Figure 4: Spearman correlation heatmap in the sentiment annotations.

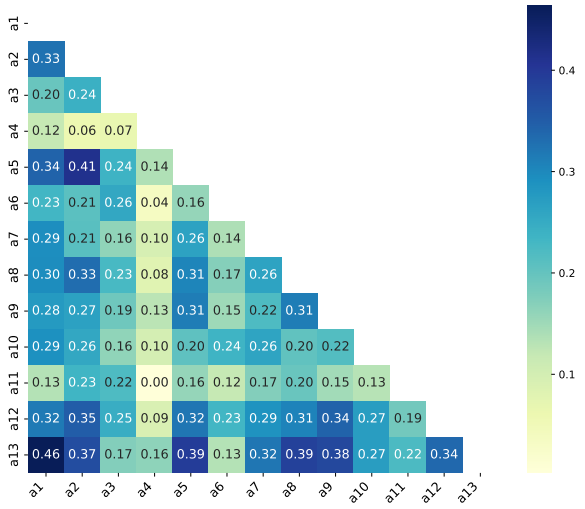

Figure 5: Cohens correlation heatmap in the sentiment annotations.

A Annotation Guidelines and Labels

In this section you can find the annotation guidelines used for the pilot and main annotation experiments (Table 7), as well as the emotion mapping to sentiment (Table 8).

B IAA further results

This part of the appendix shows some further results on the IAA on sentiment annotation. Figure 4 you can see the results when Spearman correlation is calculated, while on Figure 5 you can see the pairwise Cohen's κ values. Finally, in Table 9 you can find the distribution of the ordinal labels after the transformation as described in §4.

Sentiment	Emotions
Positive	Satisfaction, Gratitude, Joy, Love, Relief, Pleasure, Trust, Optimism, Admiration, Enthusiasm, Inspiration / Motivation, Determination, Calmness, Hope, Willingness
Neutral	Absence, Alertness / Vigilance, Nostalgia, Desire, Surprise / Shock, Anticipation / Impatience, Acceptance
Negative	Irony, Contempt, Complaint, Worry, Resentment, Despair, Anger, Sadness, Pain, Hatred, Fear / Terror, Secrecy / Secretiveness, Scorn, Suspicion, Confusion / Nervousness, Foolishness, Annoyance, Disappointment, Inadequacy, Tension, Jealousy, Bitterness, Embarrassment, Shame, Malice

Table 7: Emotion results from free-text annotation mapped to sentiment

Section	Content
Objective	The goal of this annotation task is to analyze Greek proverbs by tagging them with key linguistic and cultural features. We will explore their sentiment and emotional tone. Answer the following questions by filling the excel sheet. Do not reveal any of your personal information. The study is anonymous and voluntary so you can opt out any time.
Preliminary Question	Have you seen the proverb before?
Emotion	**What emotion or core idea does the proverb express?** Choose one or more from the list below: Absence, Irony, Contempt, Complaint, Worry, Resentment, Alertness / Vigilance, Satisfaction, Despair, Gratitude, Joy, Nostalgia, Anger, Desire, Sadness, Pain, Hatred, Fear / Terror, Love, Relief, Secrecy / Secretiveness, Pleasure, Scorn, Suspicion, Confusion / Nervousness, Foolishness, Trust, Optimism, Annoyance, Surprise / Shock, Admiration, Enthusiasm, Disappointment, Inadequacy, Anticipation / Impatience, Tension, Inspiration / Motivation, Determination, Jealousy, Bitterness, Embarrassment, Shame, Calmness, Hope, Malice, Acceptance, Willingness
Sentiment	**What sentiment does the proverb express?** Positive / Negative / Ambiguous
Example	Proverb: "The tongue has no bones, yet it breaks bones." Emotion: Bitterness, Alertness / Vigilance, Worry Sentiment: Negative
Pilot guidelines	What emotion does this proverb express? Use one emotion of your own choice in one word, using the free-text space. Do not reveal any of your personal information. The study is anonymous and voluntary so you can opt out any time.

Table 8: Annotation Guidelines for Greek Proverbs: Instructions, Emotions, Sentiment, and Example.

Ordinal Sentiment Label	Count
Mostly Ambiguous	91
Strongly Negative	83
Mostly Negative	63
Strongly Ambiguous	36
Mostly Positive	20
Strongly Positive	7

Table 9: Distribution of ordinal sentiment labels across proverbs.

C Prompts

In this section, you can find the prompts used for the model experiments as described in §3.

D Edit distance mapping

In this section, you can find an example of how Levehnstein distance was used to get the mapped localized proverb instances from the standard ones.

E Gold Maps

Here you can find the gold map of the 430 localized proverbs.

F Models and Budget

The open-sourced models were used in their quantized forms through hugging face, with only prompt configurations adjusted to optimize task performance. Llama 70B and GPT-4o were accessed respectively through AWS Bedrock and the OpenAI API. We used fixed inference settings, such as, temperature=0.7, topP=0.9. The budget reflected the

Method	Prompt
Simple Zero-shot (Z_0)	Classify the sentiment of the given proverb as Negative, Positive, or Ambiguous.
Batch Prompting (Z_b)	You are given [batch number] proverbs. Classify each proverb strictly as one of: Positive, Negative, Ambiguous. Return the results ONLY in this exact format (no explanations, no extra text): Positive, Negative, Ambiguous
Percentage Prediction Prompting (Z_p)	Estimate the percentage of how you would classify the given proverb as Positive, Negative, or Ambiguous. Percentages must sum to 100. Return strictly in this format: Positive: XX% Negative: XX% Ambiguous: XX%
Random Pick (RP)	Considering the example(s), classify the sentiment of the given proverb as Negative, Positive, or Ambiguous.
Low-Variance Selection (LVS)	(same as above)
Dynamic (DYN)	(same as above)

Table 10: Overview of prompting methods and their corresponding examples or configurations.

Standard Proverb	Localized Proverb (diff-highlighted)	Place
Άνθρωπος από γενιά και σκύλος από μάντρα	Άνθρωπος από ενιά και σκύλλος από μάντρα	Naxos
Άνθρωπος από γενιά και σκύλος από μάντρα	Τον άνθρωπον απού γενιά ΄γκαι σκύλλον απού μάντρα	Rhodes
Άνθρωπος από γενιά και σκύλος από μάντρα	Άνθρωπ΄ος από γένος και σκύλ΄ από τη μάνδρα	Thrace

Table 11: Variations of the proverb "Άνθρωπος από γενιά και σκύλος από μάντρα" with character-level differences highlighted (red = substitutions, green = insertions, red background = deletions).

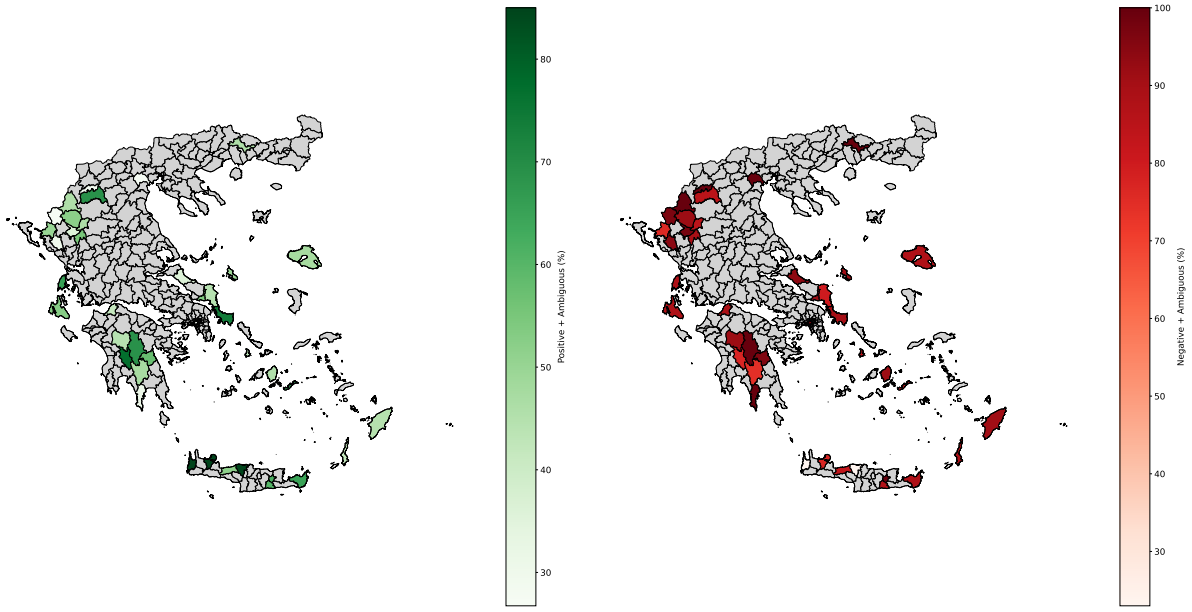

Figure 6: Sentiment maps of Greece based on the gold annotations. The green map represents the positive percentages while the red the negative.

usage costs of the APIs charged by each provider (AWS and OpenAI), with no additional expenses for model training or fine-tuning.