Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

From Monarchy to Democracy: An Analysis of LLM sentiment annotations in Albanian Political Discourse

Ueda Qorrasi[1, 2], Nathalie Pernelle [1], Aude Grezka [1]
(1) LIPN, CNRS UMR 7030, Université Sorbonne Paris Nord, 93430 Villetaneuse, France
(2) Université de Tirana, Tiranë 1010, Albanie
qorrasi@lipn.univ-paris13.fr, pernelle@lipn.univ-paris13.fr,
grezka@lipn.univ-paris13.fr

RÉSUMÉ

De la monarchie à la démocratie : une analyse des annotations de sentiments des LLM dans le discours politique albanais.
Cet article propose une analyse diachronique de l'annotation des sentiments dans le discours parlementaire albanais, en comparant un gold standard humain aux prédictions de modèles de langage de grande taille (LLMs). Deux corpus de 1 000 phrases chacun ont été constitués pour les périodes 1937–1938 et 2024–2025. Le corpus ancien reflète une variété dialectale non standardisée, tandis que le corpus contemporain est rédigé en albanais standard. Trois annotateurs humains ont évalué chaque phrase selon la polarité (positif, négatif, neutre), permettant d'établir une référence. Les mêmes données ont ensuite été annotées par ChatGPT et Gemini en configuration zero-shot, sans entraînement préalable. L'évaluation montre une accuracy qui est marginalement plus élevée pour les textes contemporains, pour les deux LLMs (accuracy variant de 75,40% à 79,50%). Cependant, l'étude montre que les caractéristiques des phrases mal annotées varient fortement selon les périodes.

ABSTRACT

This paper presents a diachronic study of sentiment annotation in Albanian parliamentary discourse, comparing human gold-standard annotations with predictions produced by Large Language Models (LLMs). Two manually curated datasets of 1,000 sentences each were compiled from different historical periods, 1937–1938 and 2024–2025. The historical corpus reflects a dialectal and non-standardized variety of Albanian, while the contemporary corpus is written in standard Albanian. All sentences were annotated for sentiment polarity (positive, negative, neutral) by three human annotators, and a gold standard was established. The same data were subsequently annotated by two LLMs (ChatGPT and Gemini) in a zero-shot setting, without any model training or fine-tuning. The evaluation shows an accuracy that is marginally higher for contemporary texts for both LLMs (accuracy ranging from 75.40% to 79.50%). However, the study shows that the characteristics of misannotated sentences vary significantly across time periods.

MOTS-CLÉS : analyse de sentiment, langue peu dotée, LLM, discours parlementaire, discours politique.

KEYWORDS: sentiment analysis, low-resource language, large language models, political discourse
.

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 165–180.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Sentiment analysis has become a widely used technique for studying political discourse, public opinion, and media communication, but this research has largely focused on high-resource languages, leaving low-resource languages with limited annotated data, tools, and reliable models (Koto et al., 2024). While extensive research exists for high-resource languages and contemporary datasets, significantly less attention has been paid to low-resource languages and historical texts. This imbalance limits our understanding of how sentiment is expressed and evolves over time in less technologically developed linguistic contexts.

Albanian represents a particularly challenging case for sentiment analysis. In addition to its low-resource status, the language exhibits substantial diachronic variation, including orthographic inconsistency, dialectal forms, and lexical change, especially in texts produced before language standardization (Riverin-Coutlée et al., 2024) . Parliamentary discourse offers a valuable source for diachronic analysis, as it reflects institutional language use and socio-political change across time. The Albanian political context offers a unique opportunity to study such evolution. Between 1937 and 2025, Albania underwent profound political transformations : from King Zog's authoritarian monarchy (1928-1939), through Italian and German occupation during World War II, to Enver Hoxha's communist dictatorship (1944-1991), and finally to post-communist democracy (1991-present) (Fischer and Schmitt, 2022) . These shifts fundamentally altered how political actors could speak, what they could say, and how sentiment could be expressed in public discourse. From a linguistic perspective, Albanian is an Indo-European language spoken by approximately 7–8 million speakers, mainly in Albania, Kosovo, North Macedonia, Montenegro, and the diaspora. It comprises two major dialect groups, Gheg in the north and Tosk in the south. During 1937–1938, Albanian had not yet been fully standardized, and official parliamentary texts often reflected dialectal and non-standard forms, particularly Gheg features. The process of language standardization was completed later, most notably with the Orthography Congress of 1972, which established a unified standard largely based on Tosk. Compared to contemporary standard Albanian, earlier varieties differ in vocabulary, morphology, syntax, and orthography, posing additional challenges for linguistic analysis and automatic processing.

This study aims to evaluate whether modern LLMs can approximate human sentiment judgments in Albanian parliamentary texts from two distant historical periods. Unlike most sentiment analysis studies, no model training is performed ; instead, LLMs are evaluated directly against a human-annotated gold standard. The central research questions are : To what extent do LLM sentiment annotations align with human gold standards in Albanian ? Does diachronic and dialectal variation affect LLM performance ? What are the linguistic challenges that can affect the annotations proposed by the LLMs ?

2 Related Work

We organize the prior literature along four axes directly relevant to this study : diachronic NLP and sentiment analysis, LLMs as annotators for low-resource languages, Albanian datasets and computational political discourse analysis.

Diachronic NLP and sentiment analysis : Research on diachronic lexical semantics has demonstrated that word meaning evolves in systematic ways over time, posing significant challenges for the automatic analysis of historical texts. Many approaches rely on diachronic word embeddings to detect and characterize semantic change in English (Hamilton et al., 2016), but also in low-resource lan-

guages such as Croatian (Dukić et al., 2025). These approaches specifically examine how neighboring embeddings evolve over time. There is no study of this kind for Albanian, but one cannot assume that words expressed in pre-standard Albanian retain the same meaning as their modern equivalents. More recent work has extended this line of research to contextualized language models : HistBERT, a BERT-based model pre-trained on historical corpora, demonstrates that models trained exclusively on contemporary data fail to capture historical semantic patterns, while temporally adapted pre-training yields representations that better reflect diachronic variation (Wenjun Qiu and Xu, 2022). Research in historical sentiment and emotion analysis emphasizes that models trained on contemporary language often fail to generalize to earlier periods due to shifts in emotional expression, orthography, and discourse conventions (Hellrich et al., 2019) (Kryeziu, 2018).

LLMs as Annotators for low-resource languages : A growing body of research investigates the use of LLMs as annotators for NLP tasks. While LLM-based annotation offers scalability and reduced costs, recent studies focusing on low-resource languages report notable limitations. In such settings, LLMs often produce inconsistent annotations, exhibit bias toward high-resource language patterns, and underperform traditional supervised or human-annotated baselines for tasks such as sentiment annotation (Jadhav et al., 2025). These findings suggest that although LLMs can support data generation, their outputs require careful validation, particularly in low-resource and culturally specific contexts. Furthermore, recent studies have investigated how prompt engineering strategies influence the performance of LLMs on sentiment-related tasks or irony detection : carefully designed few-shot prompts, reasoning strategies, and instruction tuning can significantly improve LLM performance compared to naive zero-shot settings (Kutuzov et al., 2018) (Schmitt et al., 2026). These findings suggest that while LLMs possess strong latent capabilities for sentiment understanding, their effectiveness on nuanced tasks such as irony detection is highly sensitive to prompt formulation.

Albanian corpus for sentiment analysis : For Albanian NLP, available sentiment resources remain extremely limited. The AlbMoRe corpus (Çano, 2023) constitutes one of the first sentiment-annotated datasets for Albanian, providing 800 labeled movie reviews. In addition, the EduSenti (Nuci et al., 2024) dataset offers further sentiment-labeled data for Albanian, contributing to the development and evaluation of sentiment classification models in low-resource settings. Building on such resources, recent work has explored low-resource fine-tuning of transformer-based models, such as RoBERTa, demonstrating that pretrained models can achieve promising results for Albanian sentiment classification despite limited labeled data. However, both the available datasets and modeling approaches are restricted to contemporary text and do not account for diachronic variation or institutional and political discourse.

Political discourse analysis : Research in computational political discourse analysis shows that sentiment and stance detection models can reveal longitudinal patterns in ideological communication and public rhetoric, but require careful adaptation to domain- and time-specific language use (Abercrombie and Batista-Navarro, 2022). Studies analyzing large corpora of parliamentary speeches have used topic and sentiment classification to investigate patterns across political parties and over time, showing that automated models can capture meaningful trends in political discourse and provide insights into party-specific communication and topical emphasis (Pätz et al., 2025). In a complementary line of work, the AgoraSpeech dataset offers a high-quality corpus of election speeches annotated for multiple NLP tasks, including sentiment analysis and polarization detection (Sermpezis et al., 2025).

Low-resource languages such as Albanian remain largely underexplored, particularly in the context of historical sentiment dynamics and semantic evolution. The intersection of diachronic semantic modeling and sentiment analysis remains an open challenge, to which this study contributes. We

think that it can be important to establishing a baseline for future fine-tuned models, to evaluate general-purpose LLMs in a zero-shot setting, and analyse the linguistics difficulties that can lead to erroneous annotations.

3 Construction of a Modern and a Historical Dataset

The objective of this corpus is to provide a politically annotated resource covering two periods that are both historically distant and politically contrasting, enabling the development and evaluation of sentiment analysis approaches for Albanian across radically different institutional and linguistic contexts.

3.1 Gold Standard Construction

Each dataset comprises 1,000 sentences drawn from Albanian parliamentary and political discourse, a size consistent with comparable low-resource annotation efforts (Çano, 2023) (Nuci et al., 2024) and sufficient for reliable inter-annotator agreement computation. In what follows, we adopt the terms Historic Dataset (HD) for the 1937–1938 corpus and Modern Dataset (MD) for the 2024–2025 corpus. The HD was extracted from official records of the Albanian Parliament and reflects pre-standardization Albanian, characterized by Gheg dialectal forms, archaic vocabulary, and non-standard syntactic constructions. For the historic period, this size also reflects the limited availability of digitized parliamentary records from 1937–1938. The MD was collected primarily from the official website of the Albanian Parliament, supplemented by Top Channel (top-channel.tv), a major Albanian online news outlet, both of which represent contemporary standard Albanian. Sentence selection followed a purposive sampling strategy covering legislative discussion, procedural statements, and evaluative political discourse, without imposing any target sentiment distribution.Fragments, headings, and purely procedural expressions were excluded, only complete, self-contained sentences were retained. The two corpora differ substantially in their linguistic profiles. The HD reflects pre-standardization Gheg Albanian, with dialect-specific vocabulary and non-standard morphology, while the MD adheres to contemporary standard Albanian norms. Foreign-language borrowings are present in both but differ in origin : French and Turkish terms dominate the historic corpus (e.g. budget, en bloc), while the modern corpus shows higher frequencies of English and Italian borrowings related to law, governance, and European institutions. Sentiment annotation was carried out by three native speakers of Albanian with complementary academic backgrounds : two trained linguists specializing in discourse analysis and Albanian language variation, and one historian with expertise in Albanian political and institutional history. Each sentence was independently annotated by all three annotators following a detailed annotation guide defining three sentiment categories. Positive sentiment covers expressions of approval, praise, or favorable judgment. Negative sentiment covers criticism, disapproval, or unfavorable judgment. Neutral sentiment covers procedural or descriptive discourse lacking explicit evaluative stance. Annotators were instructed to consider historical and socio-political context in cases of irony, implicit evaluation, or ambiguous rhetorical convention, and consulted dialectal and historical dictionaries when interpreting unfamiliar Gheg forms. Disagreements were more frequent in the historic corpus due to dialectal complexity and were resolved through collective discussion to produce a single gold-standard label per sentence. Full annotation guidelines are provided in Appendix A. Following

the independent annotation phase, all cases of disagreement were reviewed and resolved to produce a single gold-standard label for each sentence. The majority label was adopted where two of three annotators agreed ; remaining cases were discussed until consensus was reached through linguistic and contextual argumentation.

3.2 Descriptive Corpus Statistics and Evaluation Metrics

Descriptive statistics were computed for both datasets to characterize their linguistic profiles and contextualize the sentiment distribution (see Table 1).

The HD is characterized by a substantially higher average sentence length than the MD, indicating more complex syntactic constructions and extended formulations in parliamentary discourse. The HD contains more negative and less neutral sentiment, while the MD shows a higher proportion of neutral language. This supports the political hypothesis outlined earlier : in democratic settings, parliamentary discourse becomes more institutional, procedural, and legally framed, leading to more neutral coding. The higher negativity in the HD may reflect the formal space for opposition under the monarchy, expressed through deferential and indirect rhetoric rather than the direct criticism typical of contemporary discourse.

Measure	**Historic Dataset**	**Modern Dataset**
Total word count	24,345	19,502
Average sentence length (words)	24.3	19.5
Average word length (characters)	4.61	4.77
Negative sentiment (%)	39.5	36.4
Neutral sentiment (%)	28.8	34
Positive sentiment (%)	31.7	29.6

TABLE 1 – Descriptive statistics and sentiment distribution across datasets

3.3 Inter-Annotator Agreement

Inter-annotator agreement was consistently high across both datasets. For the HD, Cohen's Kappa scores ranged from 0.7565 to 0.8559, and for the MD, from 0.7534 to 0.8705, both indicating substantial agreement. Per-label scores, reveal a consistent pattern across datasets : negative sentiment achieved the highest agreement (HD : k = 0.839 ; MD : k = 0.845), followed by positive sentiment (HD : k = 0.834 ; MD : k = 0.812), while neutral sentiment yielded the lowest, though still substantial, scores (HD : k = 0.750 ; MD : k = 0.756). While the language of this corpus is standardized and more accessible, disagreement persisted in cases involving implicit evaluation or neutrality, particularly in procedurally oriented parliamentary statements. These agreement scores confirm the overall reliability of the annotation process and support the use of the resulting gold standards for subsequent evaluation.

3.4 Lexical and Sentiment Vocabulary Analysis

A comparative lexical analysis was conducted to characterize diachronic differences in evaluative language across the two corpora. Analysis focused on three dimensions : sentiment-bearing vocabulary, institutional terminology, and N-gram patterns reflecting discourse conventions of each period. The shift in institutional vocabulary between the two periods reflects the broader transformation of the Albanian state. The HD is dominated by formal and archaic institutional vocabulary reflecting monarchical parliamentary discourse, with terms such as, ligja ("law"), and mretnore ("royal") franga ari ("golden francs") . In contrast, the MD shows a standardized and contemporary political lexicon, including gjykata ("court"), KQZ ("Central Election Commission"), SPAK ("Special Anti-Corruption Structure"), and kushtetuese ("constitutional").

4 LLM performance

We have compared the results obtained by LLMs against the human-annotated gold standard. Given the diachronic and low-resource nature of the data, particular attention was paid to annotation reliability, linguistic variation, and corpus comparability across periods. LLM performance was evaluated against the gold standard using accuracy and Cohen's Kappa, with confusion matrices computed to characterize the pattern of classification errors. A stability analysis was conducted by re-annotating the sentences on which both models initially failed, each three additional times, to distinguish systematic errors from stochastic variance.
To situate LLM performance within a broader methodological context, we include an SVM classifier trained on TF-IDF features as a baseline.The model was evaluated through 10-fold stratified cross-validation across 1,000 annotated sentences per dataset. This ensures every sentence is tested on unseen data, guaranteeing reliable evaluation. Rather than competing with LLMs, the SVM serves as a reference point illustrating the gap between traditional supervised methods and zero-shot LLM annotation in a low-resource setting where labeled training data is scarce. As shown in Table 2, this gap is substantial across both datasets.

4.1 LLM Annotation Protocol

Both LLMs were evaluated in a zero-shot setting using the following standardized English prompt template. ChatGPT(GPT-5 mini, December 2025) and Gemini(Gemini 3 Flash,December 2025) were selected because they are two of the most widely used LLMs, both capable of processing multiple languages. Using two different models rather than one allows us to verify that the results reflect genuine linguistic challenges in the data, rather than limitations specific to a single model. Question : *What is the sentiment of the following sentence : positive, negative or neutral ? Give me only a single answer without any explanations.* No historical context, dialect information, or label definitions were provided to the models. Each sentence was submitted independently, without prior conversation context. For the historic dataset, the Gheg-dialect Albanian text was passed directly without transliteration or modernisation, allowing us to evaluate the models' robustness to non-standard orthography. If a response deviated from the three predefined labels (e.g., by including explanatory text), the model was prompted again using the same instructions until a valid label-only response was obtained. The prompt was formulated in English rather than Albanian because prior work has shown that multilingual models such as ChatGPT and Gemini produce more consistent

outputs when task instructions are given in English, even when the input text is in another language (Shi et al., 2022). This choice also maintains a clean separation between task specification and the Albanian input. Preliminary tests with Albanian-language prompts produced less consistent label-only responses, consistent with these prior findings.

4.2 LLM Performance Against the Gold Standard

Table 2 presents the performance of both models across both datasets. Both models achieved accuracy above 75% with Kappa values in a moderate range (0.62–0.69). These results indicate that zero-shot LLM annotation constitutes a viable starting point for low-resource sentiment tasks even without model training or fine-tuning.

A notable crossover pattern emerges : Gemini outperformed ChatGPT on the HD (76.66% vs. 75.40%), while ChatGPT outperformed Gemini on the MD (79.50% vs. 77.90%,). This suggests that the two models handle historical language variation differently, and that neither is universally superior across registers. Both models showed strong performance in identifying negative sentiment but exhibited more variability in distinguishing between positive and neutral classifications.

Dataset	Model	Accuracy (%)	Cohen's Kappa
Historic Dataset	ChatGPT	75.40	0.6298
	Gemini	76.66	0.6463
	SVM	53.90	0.2982
Modern Dataset	ChatGPT	79.50	0.6912
	Gemini	77.90	0.6659
	SVM	57.30	0.3562

TABLE 2 – Model performance across historic and modern datasets

Table 3 presents the number of correctly classified sentences per sentiment class for each model and dataset, enabling a more detailed inspection of classification patterns across categories.

DS	LLM	Positive	Negative	Neutral
MD	Gemini	189	274	316
	ChatGPT	220	288	287
HD	Gemini	239	326	201
	ChatGPT	245	308	201

TABLE 3 – Correctly classified sentences per sentiment class

To further characterize SVM performance, Table reports per-class precision, recall, and F1 score for both datasets. On the Historic Dataset, the model performed best on the negative class (F1 = 57.77%), followed by positive (F1 = 55.12%), while neutral was by far the most challenging category (F1 = 46.58%). On the Modern Dataset, performance was more balanced across classes, with positive achieving the highest F1 (58.28%), followed closely by neutral (56.93%) and negative (56.88%). The difficulty of the neutral class on historical text is consistent with findings in sentiment analysis literature, as neutral expressions tend to be more ambiguous and contextually dependent. The per-class results are detailed in Table 4.

TABLE 4 – SVM Per-Class Results on Datasets

Dataset	Class	Precision	Recall	F1-Score	Support
HD	Negative	0.5548	0.6025	0.5777	395
	Neutral	0.4980	0.4375	0.4658	288
	Positive	0.5503	0.5521	0.5512	317
MD	Negative	0.5532	0.5852	0.5688	364
	Neutral	0.5770	0.5618	0.5693	340
	Positive	0.5951	0.5709	0.5828	296

4.3 Error Analysis

To identify the sources of disagreement between the LLM predictions and the gold standard annotations, we conducted a manual error analysis using the complete set of sentences HD_{err} and MD_{err} for which both models have chosen an erroneous polarity (120 sentences from HD, 116 from MD). Error categories were organized into eight types across four broader dimensions, capturing the primary linguistic challenges faced by the LLMs. Each selected sentence have been associated to one or several categories by the human annotators. Error categorization was performed by two human annotators working independently, following the same label taxonomy described in Table 6. Disagreements between annotators were resolved through discussion. Table 5 presents the distribution of error categories across both datasets. HD_{err} is dominated by archaic language features (96 instances), reflecting the orthographic and morphological complexity of pre-standardization Albanian, followed by contextual knowledge dependencies (50 instances) and complex evaluative structures (37 instances). In contrast, the MD_{err} showed no archaic language errors but exhibited greater challenges in contextual knowledge dependencies (63 instances), complex evaluative structures (44 instances), and implicit evaluative expressions (33 instances), consistent with the politically dense and rhetorically sophisticated nature of contemporary parliamentary discourse. This distribution is compared against randomly chosen sentences where ChatGpt and Gemini both succeeded HD_{succ} and MD_{succ}. We illustrate each major error category with a concrete example drawn from the datasets, providing the Albanian source text, its English translation, the gold standard label, the LLM prediction, and an analysis of the misclassification.

	Dataset 1938 (HD)		**Dataset 2025 (MD)**	
Category	HD_{err}	HD_{succ}	MD_{err}	MD_{succ}
Archaic Language Features	96	78	0	0
Linguistic Borrowing	15	37	14	16
Implicit Evaluative Expression	18	42	32	53
Misunderstood Negation	20	3	13	9
Idiomatic Expression	20	13	6	23
Contextual Knowledge Dependencies	39	70	58	95
Complex Evaluative Structure	32	42	35	48
Other	25	8	35	9

TABLE 5 – Distribution of the categories among successful and incorrect annotations

Archaic features captures errors arising from dialectal spellings, non-standard Gheg morphology,

and obsolete vocabulary specific to pre-standardization Albanian. These dialectal variation remains challenging for NLP systems trained primarily on contemporary standardized language varieties (Joshi et al., 2024). Example :*"Duhet t'a dijne si Zoti Dr. Simonidhi, ashtru z.Fejzi Alizotti, se me nji budget prej 4-500.000 franga ari nuk nund te behen gjera e medhaja"* (Translation :"*They should know it, like Mr. Dr. Simonidhi, as well as Mr. Fejzi Alizotti, that with a budget of 400–500,000 gold francs, great things cannot be done"*, Gold standard : Negative, LLM : Neutral). Linguistic borrowing refers to loanwords from French, Turkish, English, and Italian that may introduce semantic ambiguity or domain-specific usage patterns (Winata et al., 2023). Borrowing errors were evenly distributed across both datasets (HD_{err} : 15 ; MD_{err} : 14), suggesting that this source of difficulty is period-independent. Example :*"Edhe një herë, ndjesë, kolegë, që u zgjata, por mendoj se është një informacion adekuat jo vetëm për ju, por, mbi të gjitha, për qytetarët."* (Translation : *"Once again, I apologize, colleagues, for going on at length, but I believe this is adequate information not only for you, but, above all, for the citizens."*, Gold standard : Positive, LLM : Neutral).

Pragmatic phenomena encompasses three sub-categories : irony and sarcasm detection, implicit evaluative expression, and misunderstood negation, each of which requires interpretation beyond literal lexical meaning. Irony and sarcasm are figurative language forms that routinely mislead neural sentiment models because their surface polarity contradicts their intended meaning (Joshi et al., 2017). Example : *"Si mund të vazhdohet, se ma tha një zonër na kanë thënë që ne të kujdesemi për të moshuarit, si qenka tani të zgjedhëm një të moshuar të na qeverisë"* (Translation : *"How can this go on, since a lady told me we should take care of the elderly ; so how is it now that we choose an elderly person to govern us ?"*, Gold standard : Negative, LLM : Neutral). Implicit evaluative expression refers to sentiment conveyed indirectly through discourse structure, pragmatic cues, or culturally shared assumptions rather than explicit evaluative vocabulary. Capturing such implicit signals remains an open challenge (Cui et al., 2023). Misunderstood negation encompasses structures where negative particles or affixes reverse sentiment polarity. Correctly identifying the syntactic and semantic scope of negation operators is a fundamental difficulty for models, since negation can be expressed both explicitly (e.g., not) and implicitly through meaning-reversing expressions (Sineva et al., 2021) (Petcu et al., 2025). Negation errors were more prevalent in HD_{err} (20 instances) than the MD (13 instances). Example (HD_{err}) : *"Nuk kërkojmë shumë gjana, por të pakten ata vllaëznit t'one përtej kuf'init të jetojnë si njerês."* (Translation : *"We do not ask for many things, but at least let our brothers beyond the border live as human beings."*, Gold standard : Positive , LLM : Negative). Idiomatic Expression category covers fixed or semi-fixed expressions whose evaluative meaning cannot be derived compositionally from their component words. Idiomatic errors were more frequent in HD_{err} (20 instances) than in MD_{err} (6 instances), reflecting the greater density of formulaic political and parliamentary idioms in the monarchical period. Example : *"I marrin dhe i lexojnë fije për fije dhe ne bëjmë debat sa orë që na lë Gjykata Kushtetuese."* (Translation : *"They take them and read them thread by thread, while we debate for as many hours as the Constitutional Court allows us."*, Gold standard : Negative, LLM : Neutral).

Contextual Knowledge Dependencies refer to cases where sentiment interpretation requires extra-linguistic knowledge of political actors, historical events, institutional structures, or temporal context. Recent work benchmarking LLMs on political tasks has confirmed that such contextual reasoning remains a systematic limitation of current models (Liang et al., 2026). This was the second most frequent error category in MD_{err} (44 instances) and also substantial in HD_{err} (37 instances). Example : Albanian :*"Për pakësimin e taksës për shoferë, jam dakord, po sa për taksën e benzinës nuk e pranoj."* Translation : *"Regarding the reduction of the tax for drivers, I agree, but as for the fuel tax, I do not accept it."* Gold standard : Positive, Gemini : Neutral, ChatGPT : Negative

To identify which linguistic features systematically challenge LLMs in historical Albanian sentiment analysis, we conducted a logistic regression. For each dataset, sentences where both Gemini and ChatGPT failed, i.e. HD_{err} and MD_{err} (Y=1), were compared against HD_{succ} and MD_{succ} (Y=0), using binary feature vectors that encode the presence of the 8 different error categories (cf. table 5). For HD, the model achieved an accuracy of 70.4%, identifying four significant predictors associated with correct classification : Implicit Evaluative Expression, Linguistic Borrowing, Contextual Knowledge Dependencies, and Complex Evaluative Structure. The same procedure applied to MD yielded an accuracy of 74.6%, confirming three of the four predictors : Contextual Knowledge Dependencies, Implicit Evaluative Expression, and Complex Evaluative Structure, while Idiomatic Expression emerged as an additional significant predictor. Across both datasets, Contextual Knowledge Dependencies was the strongest and most consistent predictor (odd-ratio OR=0.33 in HD, OR=0.25 in MD), indicating that sentences requiring historical and political contextual knowledge are significantly less likely to be misclassified while negation and archaic language features doesn't seem to play an important role in misclassifications (Table 6).

TABLE 6 – Logistic Regression Results HD vs. MD

	HD_{err}			MD_{err}		
Category	**Coef**	**OR**	**Sig**	**Coef**	**OR**	**Sig**
Misunderstood Negation	+1.18	3.24	ns	+0.05	1.05	ns
Archaic Language Features	+0.56	1.76	ns	–	–	–
Autre	+0.53	1.69	ns	+0.54	1.71	ns
Idiomatic Expression	+0.32	1.38	ns	−1.05	0.35	*
Complex Evaluative Struct.	−0.78	0.46	*	−0.80	0.45	*
Contextual Knowledge Dep.	−1.12	0.33	**	−1.37	0.25	***
Linguistic Borrowing	−1.15	0.32	**	−0.28	0.76	ns
Implicit Evaluative Exp.	−1.23	0.29	**	−0.70	0.50	*

Acc : 70.4% (HD), 74.6% (MD) * $p<0.05$ ** $p<0.01$ *** $p<0.001$ ns = not significant

4.4 Stability Analysis

To assess whether classification errors reflected systematic limitations or stochastic variance, we re-annotated HD_{err} and MD_{err} sentences using ChatGPT where both models initially failed (120 from HD, 116 from MD) three additional times per model. Table 7 shows that the results confirmed high error stability : 63% of errors in HD and 57-71% in MD were systematic (0/3 correct), indicating structural rather than random failures. we define systematic errors as sentences for which a model produced an incorrect label on all three re-annotation trials (0 out of 3 correct). This distinguishes failures that are stable and reproducible, reflecting genuine model limitations from occasional errors that may reflect stochastic variance in model output. Nevertheless, majority voting across three trials improved accuracy by 12-18 percentage points for both models. Notably, ensemble gains were larger in MD than HD (18% vs. 15% for ChatGPT), suggesting that contemporary discourse presents more genuine model uncertainty, whereas historical errors primarily reflect knowledge gaps.

These findings indicate that while ensemble methods offer practical improvements, addressing systematic errors requires data-level interventions such as specialized lexicons or domain-specific

training rather than multiple sampling alone.

TABLE 7 – Stability analysis of ChatGPT and Gemini on HD_{err} and MD_{err}

Metric	HD_{err}		MD_{err}	
	ChatGPT	Gemini	ChatGPT	Gemini
Total phrases	120	120	116	116
Error Distribution :				
Systematic errors (0/3)	63.3%	62.5%	70.7%	56.9%
Frequent errors (1/3)	20.8%	25.8%	11.2%	27.6%
Occasional errors (2/3)	15.0%	11.7%	18.1%	15.5%
No errors (3/3)	0.8%	0.0%	0.0%	0.0%
Average success rate	17.8%	16.4%	15.8%	19.5%
Majority Voting Impact :				
Single-trial accuracy	0.8%	0.0%	0.0%	0.0%
Majority vote accuracy	15.8%	11.7%	18.1%	15.5%
Improvement	+15.0%	+11.7%	+18.1%	+15.5%

5 Discussion

The LLMs Crossover : Hypotheses : A notable crossover pattern emerges from the results : Gemini outperforms ChatGPT on the HD, while ChatGPT outperforms Gemini on the MD, consistently across both accuracy and Kappa. We advance three hypotheses to account for this. First, training data composition : Gemini may have broader coverage of historical or less-standardized multilingual text, giving it greater robustness to archaic vocabulary and non-standard orthography, while ChatGPT may have denser coverage of contemporary standard Albanian from web sources. Second, tokenization : non-standard Gheg forms may be segmented differently across the two models' subword vocabularies, affecting how well sentiment-bearing signals are preserved. Third, cross-lingual transfer : Gemini may generalize more effectively to pre-standardization Albanian if its training included languages with comparable historical variation, such as Arabic dialects or Early Modern German. These remain hypotheses to be tested rather than confirmed conclusions. The practical implication, however, is clear : in diachronic low-resource settings, model selection should be validated separately for each time period and not assumed to transfer across registers.

Political Constraints and Sentiment Expression The transition from authoritarian monarchy to parliamentary democracy should, in principle, leave a detectable imprint on how sentiment is expressed : authoritarian environments displace criticism into deferential and implicit channels, while democratic environments normalize direct opposition. Our data offer consistent support for this hypothesis. The HD shows higher negative sentiment (39.5% vs. 36.4%) but markedly lower neutral sentiment (28.8% vs. 34.0%). This is not a retreat from evaluative engagement in the MD, but a change in its form : contemporary speakers express opposition through legally and procedurally framed language coded as neutral, while historical speakers used formulaic, affect-laden deference that annotators classified as evaluative. The lexical evidence confirms this. The HD's dominant evaluative n-grams are deferential : lutem qeverisë ('I plead to the government'), ndërshmja qeveri mretnore ('the honorable royal government'), while the MD's are institutional : gjykata kushtetuese, SPAK. Deference and legal invocation are both evaluative stances, but they operate through entirely

different registers. This shift is the linguistic signature of the regime transition our title names, and it carries a methodological implication : annotation schemes calibrated on contemporary democratic discourse may systematically undercount the evaluative content of authoritarian-era texts.

6 Conclusion and Future Work

This study demonstrates that zero-shot LLMs can serve as viable sentiment annotators for Albanian parliamentary discourse across two historically and linguistically distant periods, (accuracy 75.40–79.50%) without any task-specific training. This represents a meaningful result for a low-resource language with significant diachronic variation, and establishes a first baseline for future fine-tuned models. A logistic regression analysis further revealed that LLM errors are primarily driven by sentences requiring domain-specific historical and political knowledge, indirect or ironic sentiment, and complex evaluative structures rather than surface-level archaic forms. Contextual Knowledge Dependencies was the strongest and most consistent predictor across both datasets (OR=0.33 in HD, OR=0.25 in MD), suggesting that the inability to interpret historical and political context represents the most critical limitation of LLMs when applied to this corpus.

Several directions for future work follow from the present study. First, expanding the annotated corpora particularly for the historically complex period of communist-era discourse (1944–1991), which is conspicuously absent from the current study and would allow for a more complete picture of sentiment evolution across Albania's political transitions. It could reveal how totalitarian constraints shaped evaluative language in ways distinct from both the monarchical and contemporary democratic registers. Second, fine-tuning multilingual transformer models (e.g., XLM-RoBERTa) directly on the gold-standard annotations produced in this study could substantially close the gap between LLM zero-shot performance and human-level annotation. Third, a more fine-grained annotation scheme, distinguishing, for instance, between institutional neutrality, implicit criticism, and genuine evaluative absence would address the persistent neutral-classification difficulty observed across both human and model annotators. Finally, the prompt engineering literature (Schmitt et al., 2026) suggests that few-shot examples and chain-of-thought instructions could improve LLM performance on Albanian sentiment tasks without requiring model training, and this represents a low-cost avenue for immediate improvement.

Limitations Despite these encouraging results, several limitations should be noted. First, the size of the annotated historical corpus remains relatively small, which may constrain the generalizability of the findings. Second, although LLMs performed rather well overall, their difficulty in consistently identifying neutral sentiment suggests that subtle or context-dependent expressions remain challenging. Third, historical language variation, including archaic vocabulary and non-standard spelling, may still introduce annotation inconsistencies. Finally, the study focuses on parliamentary discourse, which may limit the applicability of the findings to other genres or types of Albanian texts.

Références

Abercrombie, G. and Batista-Navarro, R. (2022). Policy-focused stance detection in parliamentary debate speeches. *Northern European Journal of Language Technology*, 8.

Cui, J., Fukumoto, F., Wang, X., Suzuki, Y., Li, J., and Kong, W. (2023). Aspect-category enhanced learning with a neural coherence model for implicit sentiment analysis. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics : EMNLP 2023*, pages 11345–11358, Singapore. Association for Computational Linguistics.

Dukić, D., Barić, A., Čuljak, M., Jukić, J., and Tutek, M. (2025). Characterizing linguistic shifts in Croatian news via diachronic word embeddings. In Piskorski, J., Přibáň, P., Nakov, P., Yangarber, R., and Marcinczuk, M., editors, *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 108–115, Vienna, Austria. Association for Computational Linguistics.

Fischer, B. J. and Schmitt, O. J. (2022). *A Concise History of Albania*. Cambridge Concise Histories. Cambridge University Press.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hellrich, J., Buechel, S., and Hahn, U. (2019). Modeling word emotion in historical language : Quantity beats supposed stability in seed word selection. In Alex, B., Degaetano-Ortlieb, S., Kazantseva, A., Reiter, N., and Szpakowicz, S., editors, *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 1–11, Minneapolis, USA. Association for Computational Linguistics.

Jadhav, S., Shanbhag, A., Thakurdesai, A., Sinare, R., and Joshi, R. (2025). On limitations of LLM as annotator for low resource languages. In Abbas, M., Yousef, T., and Galke, L., editors, *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 277–282, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Joshi, A., Bhattacharyya, P., and Carman, M. J. (2017). Automatic sarcasm detection : A survey. *ACM Comput. Surv.*, 50(5) :73 :1–73 :22.

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., and Dippold, D. (2024). Natural language processing for dialects of a language : A survey.

Koto, F., Beck, T., Talat, Z., Gurevych, I., and Baldwin, T. (2024). Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon. In Graham, Y. and Purver, M., editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 298–320, St. Julian's, Malta. Association for Computational Linguistics.

Kryeziu, S. (2018). The path of standard albanian language formation. *European Journal of Social Science Education and Research*, 5 :106–116.

Kutuzov, A., Øvrelid, L., Szymanski, T., and Velldal, E. (2018). Diachronic word embeddings and semantic shifts : a survey. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liang, Y., Yang, L., Wang, C., Xia, C., Meng, R., Xu, X., Wang, H., Payani, A., and Shu, K. (2026). Benchmarking llms for political science : A united nations perspective.

Nuci, K. P., Landes, P., and Di Eugenio, B. (2024). RoBERTa low resource fine tuning for sentiment analysis in Albanian. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics,*

Language Resources and Evaluation (LREC-COLING 2024), pages 14146–14151, Torino, Italia. ELRA and ICCL.

Petcu, R., Bhargav, S., de Rijke, M., and Kanoulas, E. (2025). A comprehensive taxonomy of negation for NLP and neural retrievers. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics : EMNLP 2025*, pages 15511–15533, Suzhou, China. Association for Computational Linguistics.

Pätz, L., Beyer, M., Späth, J., Bohlen, L., Zschech, P., Kraus, M., and Rosenberger, J. (2025). Analyzing german parliamentary speeches : A machine learning approach for topic and sentiment classification.

Riverin-Coutlée, J., Kapia, E., and Gubian, M. (2024). Dialect change and language attitudes in albania. *Language Variation and Change*, 36(2) :219–242.

Schmitt, M., Schwerk, A., and Lempert, S. (2026). Enhancing sentiment classification and irony detection in large language models through advanced prompt engineering techniques.

Sermpezis, P., Karamanidis, S., Paraschou, E., Dimitriadis, I., Yfantidou, S., Kouskouveli, F., Troboukis, T., Kiki, K., Galanopoulos, A., and Vakali, A. (2025). Agoraspeech : A multi-annotated comprehensive dataset of political discourse through the lens of humans and AI. *CoRR*, abs/2501.06265.

Shi, F., Suzgun, M., Freitag, M., Wang, X., Srivats, S., Vosoughi, S., Chung, H. W., Tay, Y., Ruder, S., Zhou, D., Das, D., and Wei, J. (2022). Language models are multilingual chain-of-thought reasoners.

Sineva, E., Grünewald, S., Friedrich, A., and Kuhn, J. (2021). Negation-instance based evaluation of end-to-end negation resolution. In Bisazza, A. and Abend, O., editors, *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 528–543, Online. Association for Computational Linguistics.

Wenjun Qiu and Xu, Y. (2022). Histbert : A pre-trained language model for diachronic lexical semantic analysis.

Winata, G., Aji, A. F., Yong, Z. X., and Solorio, T. (2023). The decades progress on code-switching research in NLP : A systematic survey on trends and challenges. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics : ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Çano, E. (2023). Albmore : A corpus of movie reviews for sentiment analysis in albanian.

A Appendix : Sentiment Annotation Guidelines

Annotators were asked to assign one of three sentiment polarity labels : positive, negative or neutral, to each sentence drawn from Albanian parliamentary discourse.

A.1 Label definitions :

Positive :The sentence expresses approval, praise, endorsement, satisfaction, or a favorable evaluation of a person, institution, policy, or action. Positive evaluative adjectives, verbs of approval, and expressions of support are typical markers.
Example (modern) : "Kjo reformë përbën një hap të rëndësishëm drejt konsolidimit të demokracisë."

("This reform represents an important step toward the consolidation of democracy.") → Positive
Negative : The sentence expresses criticism, disapproval, reproach, dissatisfaction, or an unfavorable evaluation. Negative lexical items, verbs of condemnation, and expressions of rejection are typical markers.
Example (historic) : "Nuk mund të pranohet ky propozim i qeverisë." ("This proposal of the government cannot be accepted.") → Negative
Neutral : The sentence is procedural, descriptive, or institutional in nature, and does not carry an explicit evaluative stance. This includes statements of fact, procedural announcements, and legislative formulations without evaluative framing.
Example (modern) : "Kuvendi mblidhet në seancë plenare çdo të martë." ("The Assembly holds a plenary session every Tuesday.") → Neutral

A.2 Special cases and instructions

Irony and sarcasm : When a sentence uses positive surface language to convey a negative meaning (or vice versa), annotators should label the intended meaning rather than the literal surface form. When irony is suspected, annotators were instructed to note it and discuss with the team.
Mixed sentiment : When a sentence contains both positive and negative elements, annotators should assign the label that reflects the dominant evaluative orientation of the sentence as a whole.
Historical language : For sentences from the 1937–1938 corpus, annotators were permitted to consult Gheg dialect dictionaries and historical parliamentary records to resolve lexical ambiguities. Contemporary interpretations should not be projected onto historical expressions.
Implicit evaluation : Sentences that convey evaluation through discourse structure or pragmatic implication rather than explicit vocabulary should be labeled according to the evaluable communicated meaning in context.

Disagreement resolution : After independent annotation, all cases of disagreement were reviewed collectively. The majority label was adopted where two of three annotators agreed. Remaining cases were discussed until consensus was reached through linguistic and contextual argumentation.

A.3 Per-label inter-annotator agreement

TABLE 8 – Inter-annotator agreement per label

Dataset	Metric	Positive	Negative	Neutral
HD	Ann1 vs Ann2	0.788	0.795	0.679
	Ann1 vs Ann3	0.876	0.879	0.806
	Ann2 vs Ann3	0.839	0.842	0.764
	Average	**0.834**	**0.839**	**0.750**
MD	Ann1 vs Ann2	0.871	0.910	0.827
	Ann1 vs Ann3	0.754	0.807	0.701
	Ann2 vs Ann3	0.811	0.819	0.741
	Average	**0.812**	**0.845**	**0.756**

Table 8 presents the pairwise inter-annotator agreement scores for each sentiment label across the HD and MD datasets. The agreement between annotators was generally high for the positive and negative

labels in both datasets, indicating a consistent interpretation of clearly polarized instances. In contrast, the neutral label obtained comparatively lower agreement scores, suggesting that neutral instances were more ambiguous and difficult to distinguish consistently among annotators. Overall, these observations highlight the inherent complexity of neutral sentiment annotation and the importance of multiple annotators in ensuring annotation reliability.