

Article

A Topic-Aware Structured Semantic Representation Framework for Sentiment Analysis in Greek Social Media

Kyriakos Skoularikis *  and Ilias K. Savvas 

Department of Digital Systems, University of Thessaly, 41500 Larissa, Greece; isavvas@uth.gr

* Correspondence: kyriskoul1@uth.gr; Tel.: +30-6940403249

Abstract

Sentiment analysis for Greek social media texts remains challenging because of limited annotated resources, linguistic variation, and domain-dependent sentiment expression. This study presents a topic-aware, lexicon-guided framework for sentiment classification across five reference domains in Greek social media. Domain-specific sentiment lexicons are activated according to the relevant domain and transformed into a structured representation comprising a token-level multi-channel lexical matrix and aggregate lexical descriptors. A fusion convolutional neural network combines these complementary components to classify sentiment while retaining explicit lexical evidence for inspection. The evaluation follows a leakage-free protocol: lexicons are constructed exclusively from the sentiment inner-training subset, validation data are used for model selection, and a held-out test set is reserved for final evaluation. The proposed fusion CNN achieved the strongest held-out sentiment result among the evaluated models, with an Accuracy of 0.8029 and a Macro-F1 of 0.7883, exceeding TF-IDF + Linear SVM and fine-tuned GreekBERT baselines in the present experimental setting. Ablation results show that the token-level lexical matrix and global descriptors provide complementary information. For domain routing, GreekBERT late fusion achieved an Accuracy of 0.9162 and a Macro-F1 of 0.9116. When lexicon activation used predicted rather than reference domains, the end-to-end sentiment pipeline achieved a Macro-F1 of 0.7569. These findings indicate that explicit domain-specific lexical knowledge can support an interpretable sentiment representation while making the effects of lexical coverage and topic-routing uncertainty visible.

Keywords: sentiment analysis; topic-aware modeling; Greek social media; lexicon-based methods; structured semantic representation; convolutional neural networks; GreekBERT; text classification



Academic Editor: Wajahat Ali Khan

Received: 9 June 2026

Revised: 15 July 2026

Accepted: 23 July 2026

Published: 26 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Sentiment analysis has become a central task in natural language processing, enabling the large-scale interpretation of user-generated content across domains such as social media, e-commerce, and public discourse. Social media platforms constitute a particularly challenging environment due to their short, informal, and rapidly evolving linguistic expressions, making sentiment analysis especially demanding in morphologically rich and low-resource languages such as Greek. Despite substantial progress, accurately modeling sentiment in short, informal texts remains challenging, particularly in low-resource languages where annotated data and computational resources are limited [1,2]. These constraints further amplify the need for models that can incorporate structured linguistic knowledge beyond purely data-driven learning.

Existing approaches to sentiment analysis can be broadly categorized into lexicon-based, machine learning, and deep learning methods. Lexicon-based approaches offer high interpretability and explicit semantic knowledge but typically rely on feature aggregation, compressing token-level sentiment signals into global descriptors and consequently losing local semantic organization [1,3]. In contrast, deep learning methods, including convolutional neural networks and transformer-based architectures such as BERT, learn representations directly from data and capture contextual dependencies more effectively [4,5]. However, these approaches often rely on large annotated datasets, while their reliance on implicit latent representations limits the explicit integration of domain-specific semantic knowledge and reduces model interpretability, particularly in specialized application domains.

More recently, large language models have further advanced the state of the art in sentiment analysis through large-scale pretraining and contextual modeling [6]. While highly effective, these models primarily rely on implicit contextual representations and provide limited support for explicitly incorporating structured semantic knowledge and domain-aware lexical constraints into the representation process. Recent advances in parameter-efficient fine-tuning and quantization have substantially improved the feasibility of adapting large language models for Greek and other low-resource settings [7,8].

Recent studies have explored increasingly sophisticated mechanisms for integrating lexical knowledge with neural architectures, including gate-level lexicon–neural fusion and token-level semantic modulation [9,10]. These approaches enhance neural representation learning by incorporating lexical knowledge within pretrained language models or recurrent neural architectures. However, they primarily improve internal neural representations after textual encoding, while lexical knowledge typically remains an auxiliary component of the learning process. Comparatively less attention has been devoted to constructing structured semantic representations prior to sentiment classification through the dynamic activation of domain-specific lexical knowledge. Such topic-aware structured semantic representations preserve explicit token-level semantic information, maintain local semantic organization through dynamically activated domain-specific lexicons, and provide an interpretable semantic input space for downstream representation learning.

To address this gap, we propose an interpretable knowledge-guided framework that complements, rather than replaces, contemporary large language models through a lexicon-guided structured semantic representation for sentiment classification. Rather than aggregating lexical information into global sentiment descriptors, the proposed framework transforms domain-specific lexical knowledge into an explicit semantic representation in which each token is encoded through a set of sentiment-aware attributes. The resulting structured semantic representation preserves token-level information and local semantic interactions while providing an interpretable semantic input space for downstream learning. This representation is subsequently processed using convolutional neural networks to identify localized sentiment patterns within a lexicon-guided semantic feature space [4].

The proposed framework further incorporates a topic-aware mechanism in which a thematic domain determines the dynamic activation of the corresponding domain-specific lexicon. This design enables sentiment analysis to be performed within a contextually relevant semantic space, allowing lexical knowledge to adapt to domain-dependent variations in word polarity and semantic interpretation. For the controlled sentiment experiments presented in this study, collection-derived domain labels are used to activate the corresponding domain-specific lexicons, enabling the contribution of the proposed sentiment representation to be evaluated independently of the upstream topic detection module.

The proposed framework is evaluated on a multi-domain dataset of Greek social media texts, representing a realistic low-resource language setting. Experimental results

demonstrate that the proposed structured semantic representation consistently improves upon a strong lexicon-based baseline while preserving interpretability. Additional comparisons with conventional machine learning models and end-to-end evaluation settings further characterize the performance of the proposed representation under controlled and predicted-domain activation settings.

The main contributions of this work are as follows: (i) a lexicon-guided structured semantic representation that encodes token-level domain-specific lexical attributes in a multi-channel matrix and combines them with global lexical descriptors; (ii) a topic-aware framework that dynamically activates domain-specific lexicons through domain routing; (iii) a leakage-free experimental protocol with fixed inner-training, validation, and held-out test partitions for lexicon reconstruction and model selection; and (iv) an empirical evaluation including classical and transformer baselines, component ablations, domain-wise analysis, and reference-domain versus predicted-domain evaluation.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Sections 3–6 present the proposed framework, including its overall architecture, domain-specific lexicon construction, topic detection module, and knowledge-guided topic-aware sentiment analysis model. Section 7 describes the experimental setup, while Section 8 presents and discusses the experimental results. Finally, Sections 9–11 summarize the contributions of the study, discuss its limitations, and outline directions for future research.

2. Related Work

2.1. Lexicon-Based Sentiment Analysis

Lexicon-based approaches constitute one of the earliest and most interpretable paradigms in sentiment analysis, relying on predefined sentiment dictionaries to assign polarity scores to textual units and aggregate them at the document level [3,11]. Their main advantage lies in their independence from labeled data, making them particularly suitable for low-resource languages and domains with limited annotation availability. However, their effectiveness is strongly dependent on lexicon quality and domain relevance, as general-purpose lexicons often fail to capture context-dependent sentiment variations [12]. Moreover, most lexicon-based methods rely on aggregated representations, where token-level sentiment signals are reduced to a small set of global features, leading to a loss of local semantic interactions and compositional effects within the text [3]. This limitation highlights the need for approaches that preserve fine-grained lexical information within the representation space rather than compressing it [13].

2.2. Deep Learning for Sentiment Classification

Deep learning approaches have significantly advanced sentiment analysis by enabling models to learn hierarchical representations directly from textual data. Convolutional neural networks have demonstrated strong performance in sentence-level classification tasks by capturing local semantic patterns through convolutional filters [4,14], while recurrent and transformer-based architectures further improve contextual modeling by capturing sequential and long-range dependencies [5]. Transformer-based models such as BERT have become dominant due to their ability to produce contextualized embeddings that encode both semantic and syntactic information [5]. More recently, large language models have further advanced contextual language understanding and have been successfully adapted to low-resource Greek language tasks through parameter-efficient fine-tuning and quantization strategies [7,8]. Nevertheless, these approaches remain primarily focused on learning implicit contextual representations and provide limited mechanisms for explicitly incorporating structured domain knowledge, such as domain-specific sentiment lexicons, into the representation process. Consequently, their learned semantic representations are

often less interpretable than approaches that explicitly encode lexical knowledge within the representation itself.

2.3. *Lexicon–Neural Integration Approaches*

To bridge the gap between interpretability and predictive performance, lexicon–neural integration approaches have been proposed that combine lexical knowledge with machine learning and deep learning architectures. A common strategy involves augmenting feature-based representations with lexicon-derived signals, improving classification performance by injecting domain knowledge into the learning process [2,15]. More advanced methods have explored sentiment-aware embeddings and feature fusion techniques to integrate lexical and distributional information [16]. Subsequent research has explored increasingly tighter integration between lexical resources and neural architectures, moving beyond simple feature augmentation toward representation-level and model-level knowledge integration. FEAM introduces a dynamic prompting strategy combined with hierarchical convolutional attention to adapt lexical and emotional cues during contextual representation learning [10]. Similarly, SAGU-Net proposes a gate-level lexicon–neural fusion mechanism that injects sentiment lexicon information directly into recurrent neural representations through sentiment-aware gated units [9]. These studies demonstrate that explicit lexical knowledge can effectively enhance neural sentiment modeling when tightly coupled with contextual representation learning.

Despite these advances, lexical knowledge is primarily integrated as an auxiliary component within the neural learning process after textual encoding. Comparatively less attention has been devoted to constructing explicit structured semantic representations before classification, where lexical knowledge itself constitutes the primary representation space rather than an additional signal provided to the classifier. This architectural distinction motivates the proposed knowledge-guided framework, which transforms topic-specific lexical knowledge into an interpretable structured semantic representation prior to downstream neural processing.

2.4. *Topic-Aware Sentiment Analysis*

The interpretation of sentiment is often shaped by thematic context, particularly in social media environments where identical lexical expressions may convey different sentiment depending on the discussion domain. This challenge is especially pronounced in Greek social media, where short, informal texts, domain-specific vocabulary, abbreviations, and rapidly evolving language increase semantic ambiguity and complicate sentiment interpretation. Consequently, topic-aware sentiment analysis has emerged as an effective strategy for incorporating contextual information into sentiment prediction.

Transformer-based models have been widely used in this setting due to their ability to capture contextual dependencies and domain-specific semantics [5,17]. Language-specific adaptations such as GreekBERT have demonstrated strong performance in Greek Natural Language Processing (NLP) tasks, including topic classification and sentiment analysis [18]. Recent studies have further demonstrated the effectiveness of transformer-based models for Greek social media analysis, including sentiment analysis on Greek Twitter data and topic classification on Greek Reddit discussions, highlighting the growing importance of contextual language models for Greek NLP applications [19,20]. Despite their strong contextual modeling capabilities, these approaches primarily rely on implicit semantic representations and rarely exploit explicit domain-specific lexical knowledge during representation construction. Consequently, although topic awareness has significantly improved sentiment classification, opportunities remain for explicitly integrating dynamically activated lexical knowledge within an interpretable structured semantic representation framework.

2.5. Research Gap and Motivation

The literature reviewed in the previous subsections demonstrates substantial progress in lexicon-based, deep learning, and topic-aware sentiment analysis. Lexicon-based methods offer transparency and explicit semantic knowledge but generally rely on aggregated document-level representations that overlook fine-grained token interactions. Deep learning and transformer-based approaches achieve strong predictive performance through contextual representation learning, while recent lexicon–neural integration methods have further strengthened the incorporation of lexical knowledge into neural architectures. Nevertheless, these approaches primarily enhance internal neural representations rather than explicitly constructing semantic representations prior to classification.

Furthermore, although topic-aware sentiment analysis has demonstrated that thematic information significantly improves contextual sentiment interpretation, relatively limited attention has been devoted to dynamically activating domain-specific lexical knowledge before representation learning. Consequently, an opportunity remains for sentiment analysis frameworks that combine topic awareness with explicit knowledge-guided semantic representations capable of preserving token-level lexical information while maintaining interpretability.

Motivated by these observations, the proposed framework introduces a knowledge-guided topic-aware structured semantic representation in which domain-specific lexical knowledge is dynamically activated according to thematic context and transformed into an explicit semantic representation prior to neural processing. The following section presents the overall architecture of the proposed framework and describes the interaction between its individual components.

3. Overall Architecture of the Proposed Framework

3.1. Framework Overview

The proposed framework follows a modular knowledge-guided architecture for topic-aware sentiment analysis in Greek social media texts. Rather than relying on a single static sentiment representation, the framework dynamically selects domain-specific semantic knowledge according to the thematic context of each input text. This design enables lexical knowledge to be incorporated directly into the representation process while preserving interpretability and token-level semantic information. The overall workflow consists of three sequential stages: (i) construction of domain-specific sentiment lexicons, (ii) topic detection and dynamic lexicon activation, and (iii) knowledge-guided structured semantic representation for sentiment classification. Figure 1 illustrates the complete architecture of the proposed framework.

3.2. Workflow Description

Given an input social media post, the framework first determines its associated thematic domain. The identified topic is subsequently used to dynamically activate the corresponding domain-specific sentiment lexicon, ensuring that lexical knowledge is interpreted within an appropriate semantic context. The activated lexicon is then employed to construct a structured semantic representation in which each token is encoded using explicit sentiment-aware lexical attributes. Finally, the resulting representation is processed by a convolutional neural network to predict the sentiment polarity of the input text.

For the experiments reported in this study, the modular design of the framework allows the topic identification and sentiment analysis components to be evaluated independently or integrated into a complete processing pipeline. To isolate the contribution of the proposed structured semantic representation, collection-derived domain labels were used to activate the corresponding domain-specific lexicons during the controlled sen-

timent experiments. Additional experiments using automatically predicted topics are presented in Section 8 to evaluate the robustness of the complete framework under realistic operating conditions.

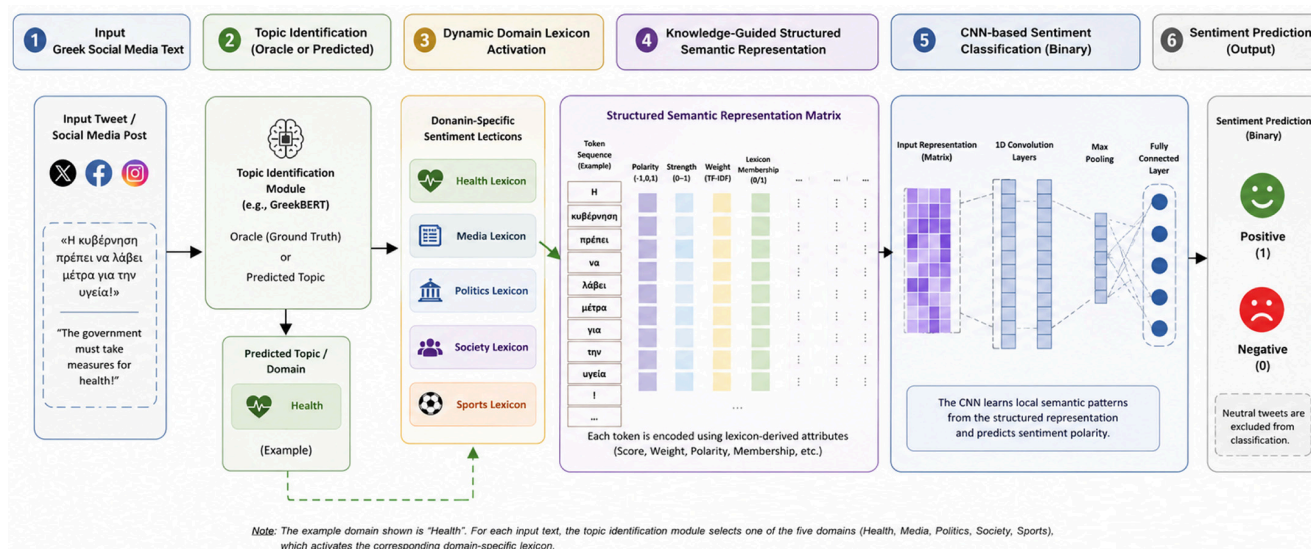


Figure 1. Architecture of the proposed framework.

4. Dataset Collection, Annotation, and Construction of Domain-Specific Lexicons

4.1. Dataset Collection and Annotation

The dataset consists of 1908 Greek-language tweets collected through the Twitter API between April and May 2023. Tweets were retrieved using a diverse set of domain-specific hashtags and query terms covering Health, Media, Politics, Society, and Sports. Each tweet was assigned a single domain label derived from the query or hashtag through which it was collected; domain labels were therefore collection-derived rather than manually annotated [21].

The distribution of tweets across the five collection-derived domains is shown in Figure 2. The domain distribution is moderately imbalanced, with Society representing the largest subset and Health the smallest. This imbalance motivates the use of macro-averaged F1-score alongside accuracy for topic-detection evaluation, so that performance is not dominated by the more frequent domains.

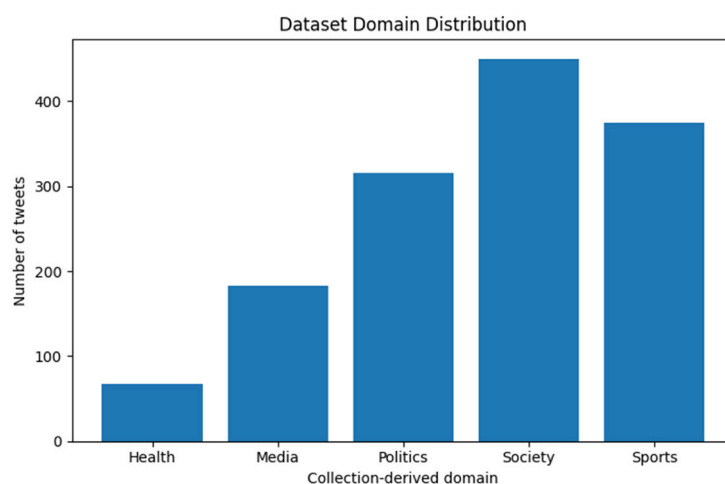


Figure 2. Distribution of the 1908 tweets across the five collection-derived domains.

During data preparation, duplicate posts, retweets, spam-like content, non-Greek tweets, and empty or excessively short posts were removed. Tweet metadata were discarded during preprocessing. Tweets containing occasional foreign terms or Greeklish expressions were retained when they remained interpretable within the Greek social media context.

Sentiment was independently assessed by four annotators using a continuous scale from -5 to $+5$. Annotators followed shared annotation guidelines and evaluated all tweets independently, without access to one another's ratings. Their assessments were used to derive a final rounded continuous sentiment label for each tweet. The final label-generation procedure was applied consistently across the full dataset before the present experiments. The archived dataset retains the final labels but not the individual annotator-level ratings; consequently, an inter-annotator agreement statistic could not be recomputed for the present revision.

This annotation procedure follows the protocol used in the authors' earlier studies on Greek social media sentiment analysis [22].

4.2. Construction of Domain-Specific Lexicons

The first stage of the proposed framework involves the construction of domain-specific sentiment lexicons tailored to Greek social media discourse. Five thematic domains are considered: Health, Media, Politics, Society, and Sports. For each domain, a separate corpus of tweets is utilized, annotated with sentiment scores in the interval $[-5, +5]$.

4.2.1. Preprocessing and Candidate Vocabulary

All corpora are processed using a standardized pipeline, including lowercasing, removal of URLs and user mentions, filtering of non-Greek tokens, stop-word removal, and normalization of diacritics. The resulting tokens form the candidate vocabulary for lexicon construction. To ensure statistical reliability, only tokens with sufficient frequency are retained.

4.2.2. Lexical Scoring Mechanism

Each candidate word is assigned a sentiment score, denoted as $Score_w$, which represents the average sentiment polarity of all sentiment-annotated tweets containing that word.

$$Score_w = \frac{1}{N_w} \sum_{i=1}^{N_w} s_i \quad (1)$$

where N_w denotes the number of sentiment-annotated tweets containing word w , and s_i represents the sentiment annotation assigned to the i -th tweet in which the word occurs.

While $Score_w$ captures the average sentiment orientation of a word, it does not reflect its discriminative importance within a domain-specific corpus. To account for this property, a second metric, $Weight_w$, is introduced:

$$Weight_w = |Score_w| \cdot TFIDF_{max}(w) \quad (2)$$

where $TFIDF_{max}(w)$ denotes the maximum Term Frequency-Inverse Document Frequency (TF-IDF) value of word w across documents within the corresponding domain corpus, reflecting its discriminative importance rather than contextual semantics.

This formulation combines sentiment intensity with domain-specific lexical importance, assigning higher weights to words that are both sentiment-bearing and highly discriminative within the corresponding thematic corpus. Compared with approaches based solely on word frequency or sentiment polarity, the proposed weighting scheme provides a more informative characterization of lexical importance for subsequent semantic representation.

4.2.3. TF-IDF-Based Filtering

To retain only informative lexical items, a filtering process based on Term Frequency–Inverse Document Frequency (TF–IDF) is applied. For each candidate word, the maximum TF–IDF value observed across all tweets within the corresponding domain corpus is used as an indicator of its discriminative importance. Although TF–IDF does not capture contextual semantics, it effectively identifies terms that are particularly informative within a specific thematic domain.

The TF–IDF criterion is used jointly with sentiment orientation and minimum occurrence requirements to retain lexical items that are both sentiment-bearing and sufficiently domain-informative. The precise eligibility criteria and the leakage-free threshold-selection procedure used in the reported experiments are specified in Section 7.2.

4.2.4. Final Lexicon Construction

A word is included in a domain-specific lexicon if it satisfies the eligibility criteria defined for the corresponding experimental setting, including minimum document frequency, non-zero sentiment score, token-length requirements, and a minimum maximum TF–IDF value. Each lexicon entry is represented by the tuple $(Score_w, TFIDF_{max}(w), Weight_w)$, providing complementary information about sentiment orientation, lexical importance, and domain specificity. The constructed lexicons constitute the domain-specific lexical resource of the proposed framework. Following topic detection, the corresponding lexicon is dynamically activated and used to construct the lexicon-guided structured representation described in Section 6. In this way, the lexicons provide the explicit lexical information that links topic routing with downstream sentiment classification.

5. Topic Detection Framework

5.1. Topic-Aware Detection Framework

In social media environments, sentiment is strongly influenced by thematic context, as the same lexical unit may exhibit different semantic and emotional roles across domains. This domain dependency is particularly pronounced in short texts such as tweets, where contextual cues are limited.

To address this challenge, the proposed framework incorporates a dedicated topic detection stage prior to sentiment analysis, enabling the dynamic activation of domain-specific lexical knowledge before semantic representation construction. Each tweet is assigned to one of five predefined domains (Health, Media, Politics, Society, and Sports), enabling the activation of domain-specific lexicons and supporting context-aware sentiment interpretation. Topic detection is formulated as a multi-class classification problem and serves as the semantic routing mechanism that dynamically selects the appropriate domain-specific lexicon for the subsequent knowledge-guided sentiment representation.

5.2. Feature Representation and Model Overview

Two complementary feature representations are investigated for topic classification. The baseline representation consists of conventional TF–IDF vectors, which capture the statistical distribution of terms within the corpus and provide a purely distributional description of tweet content [23]. The enriched representation extends this baseline by incorporating domain-specific lexicon-derived features [2], introducing explicit semantic signals that quantify the alignment between the tweet content and each domain-specific lexicon. This design enables a direct comparison between purely statistical text representations and knowledge-enriched representations, allowing the contribution of explicit lexical knowledge to topic detection performance to be systematically evaluated.

5.3. Classical Machine Learning Models

To establish baselines for the topic detection task, four classical machine learning models were evaluated: Multinomial Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Random Forest. Classical machine learning models have demonstrated strong performance in text classification tasks, particularly when combined with sparse textual representations [24].

Two feature configurations were examined. In the baseline setting, models operate solely on TF-IDF representations. In the enriched setting, the baseline TF-IDF representation is extended with five lexicon-derived features that quantify the semantic alignment of each tweet with the corresponding domain-specific lexicons.

$$X_{enrichedw} = [X_{TFIDF}, f_1, f_2, f_3, f_4, f_5] \quad (3)$$

where x_{TFIDF} denotes the conventional TF-IDF representation and $f_1, \dots, f_5$ denote the five lexicon-derived semantic features introduced in the enriched representation.

Rather than replacing the statistical text representation, these features introduce explicit semantic knowledge that complements the distributional information captured by TF-IDF. Their effectiveness depends on the ability of each classifier to exploit heterogeneous feature spaces, combining sparse textual representations with dense semantic signals. This experimental design enables a controlled evaluation of the contribution of explicit lexical knowledge across different classical machine learning classifiers.

5.4. GreekBERT-Based Topic Detection Architectures

To complement the classical machine learning approaches, a transformer-based topic detection model based on GreekBERT is also evaluated. Transformer-based architectures, and particularly language-specific adaptations such as GreekBERT, have demonstrated strong performance in Greek text classification and topic detection tasks [18–20]. In a second configuration, a late fusion strategy is employed, where the contextual embedding produced by GreekBERT is concatenated with the same 10-dimensional global lexical feature vector used throughout the proposed framework (described in Section 6.2). This feature vector summarizes tweet-level lexical information, including sentiment strength, polarity distribution, lexical coverage, and other aggregate semantic descriptors derived from the dynamically activated domain-specific lexicon, before the combined representation is passed to the final classification layer.

Overall, the two configurations enable a direct comparison between purely contextual representations and knowledge-enriched contextual representations augmented with the proposed global lexical descriptor vector, allowing the contribution of explicit lexical knowledge within transformer-based topic detection to be systematically evaluated.

5.5. Dynamic Lexicon Activation

The output of the topic detection stage is directly integrated into the proposed framework through a dynamic lexicon activation mechanism. Once a tweet is assigned to one of the predefined thematic domains, the corresponding domain-specific lexicon is automatically activated and used to construct the knowledge-guided structured semantic representation described in Section 6. This process ensures that lexical knowledge is interpreted within an appropriate thematic context, accounting for domain-dependent variations in word polarity and semantic meaning. Consequently, topic detection serves not only as a multi-class classification task but also as the semantic routing mechanism that bridges topic detection and sentiment classification within the proposed topic-aware framework.

6. Proposed Topic-Aware Sentiment Analysis Framework

The proposed framework addresses sentiment classification in Greek social media by formulating the task as structured semantic representation learning under a topic-aware and knowledge-guided setting. Rather than relying on unstructured text representations or aggregated lexical statistics alone, the model explicitly constructs a structured semantic space that encodes sentiment-related knowledge derived from domain-specific lexicons. This enables sentiment analysis to operate on explicit knowledge-guided semantic representations rather than relying solely on implicit latent representations learned directly from raw text. The overall architecture follows a multi-stage process in which topic detection precedes sentiment inference, and the predicted topic determines the activation of a domain-specific semantic lexicon. This mechanism introduces a form of semantic routing, ensuring that sentiment interpretation is conditioned on the appropriate thematic context.

6.1. Topic-Aware Lexicon Activation

The first stage of the framework performs dynamic selection of a domain-specific lexicon based on the predicted topic of each input tweet. Let t denote an input tweet and $d(t)$ denote the domain predicted by the topic detection module:

$$d(t) \in \{\text{Health, Media, Politics, Society, Sports}\} \quad (4)$$

the corresponding domain-specific lexicon L_d is subsequently activated and used for structured semantic representation construction.

This process can be interpreted as a context-dependent knowledge-guided semantic grounding mechanism, where lexical knowledge is selectively activated according to thematic context. This allows sentiment-bearing expressions to be interpreted within domain-specific semantic constraints, addressing ambiguity arising from cross-domain lexical variation.

6.2. Structured Semantic Representation Construction

Given an input token sequence $T = (w_1, w_2, \dots, w_n)$ and the activated lexicon L_d , each token is mapped into a structured semantic feature vector:

$$x_i = \phi(w_i, L_d) \quad (5)$$

where w_i denotes the i -th token of the input sequence and $\phi(\cdot)$ is the knowledge-guided mapping function that transforms lexical information into structured semantic features.

This mapping encodes explicit lexical semantic attributes, including sentiment polarity, sentiment intensity, lexical importance ($Weight_w$), and binary lexicon membership. Tokens not present in the active lexicon are represented using null semantic vectors, preserving structural alignment while distinguishing informative from non-informative elements.

The resulting representation is formalized as a structured semantic matrix:

$$X \in \mathbb{R}^{L \times C} \quad (6)$$

where X denotes the structured semantic representation matrix constructed for the input tweet, L is the maximum input sequence length and C corresponds to the number of semantic feature channels composing each token representation.

This matrix constitutes the proposed knowledge-guided structured semantic representation, in which rows preserve token order while columns encode complementary lexical semantic attributes derived from the activated domain-specific lexicon. Unlike aggregation-

based approaches, this representation preserves local semantic interactions and enables the explicit incorporation of topic-conditioned lexical knowledge before neural processing.

6.3. Semantic Pattern Modeling and Representation Fusion

A convolutional neural network is applied to the proposed knowledge-guided structured semantic representation in order to learn local semantic interaction patterns among neighboring tokens. Rather than operating directly on raw textual input, the network processes explicit semantic feature channels constructed from the activated domain-specific lexicon. The learned representation is defined as:

$$h_t = \text{CNN}(X_t) \quad (7)$$

where X_t denotes the structured semantic representation of the tweet t , $\text{CNN}(\cdot)$ represents the convolutional neural network encoder, and h_t is the resulting latent semantic representation learned from the structured semantic input.

In parallel, a global lexical descriptor, denoted by f_t , is computed from aggregated lexical statistics, summarizing corpus-level lexical information including overall sentiment intensity, lexical coverage, and polarity distribution.

$$z_t = [h_t; f_t] \quad (8)$$

where z_t denotes the final fused representation used for sentiment classification. The final sentiment prediction is obtained by applying a fully connected output layer to the fused representation z_t . For the binary sentiment task, the output logit is transformed into a probability using a sigmoid activation.

This fusion integrates two complementary levels of semantic information: (i) structured local semantic interactions derived from lexicon-informed representation learning, and (ii) global sentiment distributional properties derived from aggregated lexical signals, producing the final fused representation z_t , which serves as the input to the sentiment classifier.

6.4. Conceptual Interpretation of the Framework

Unlike conventional text classification pipelines that learn semantic representations implicitly from raw textual input, the proposed framework explicitly constructs a structured semantic representation before neural processing. This design separates knowledge representation construction from representation learning: domain-specific lexical knowledge is first organized according to thematic context and subsequently processed by the neural model. Consequently, the convolutional neural network focuses on learning interactions among explicit semantic attributes rather than discovering lexical semantics from limited training data. This separation improves interpretability, facilitates the incorporation of domain knowledge, and is particularly advantageous for low-resource settings where purely data-driven representation learning is often constrained by limited annotated data.

7. Experimental Setup

This section describes the experimental protocol used to evaluate the proposed topic-aware sentiment analysis framework. All experiments were designed to prevent leakage of sentiment-label information from validation and held-out test instances into lexicon construction, threshold selection, feature engineering, model selection, or final evaluation.

7.1. Dataset Splits and Evaluation Protocol

The dataset comprises 1908 Greek-language tweets spanning five collection-derived thematic domains: Health, Media, Politics, Society, and Sports. Each tweet is associated with a final rounded sentiment score ranging from -5 to $+5$ and a collection-derived domain label, as described in Section 4.1.

The topic detection task was formulated as a five-class classification problem. The original dataset split contained 1526 tweets in the training pool and 382 tweets in the held-out test set. To support validation-based model selection, the training pool was further divided into an inner-training subset of 1414 tweets and a validation subset of 112 tweets. Topic classifiers were trained exclusively on the inner-training subset, while the validation subset was used for model selection and early stopping. The held-out test set of 382 tweets was not used during topic-model training, validation-based selection, or any sentiment-label-dependent lexicon construction.

Sentiment classification was formulated as a binary positive/negative classification task. Tweets with neutral sentiment scores were excluded from this task. The resulting sentiment dataset was divided into 1004 inner-training tweets, 112 validation tweets, and 274 held-out test tweets. The same fixed split was used consistently across all sentiment baselines, ablation experiments, and proposed-model evaluations.

Accuracy and macro-averaged F1-score (Macro-F1) were used as the primary evaluation metrics. Macro-F1 was included to account for class imbalance and to provide balanced performance estimates across classes. Confusion matrices and class-level precision, recall, and F1-score were also examined for the final selected models.

7.2. Leakage-Free Construction of Domain-Specific Lexicons

Domain-specific sentiment lexicons were constructed exclusively from the 1004 tweets of the sentiment inner-training subset. Neither validation tweets nor held-out test tweets were used during lexicon construction. This protocol ensures that lexical sentiment scores and lexical weights are derived without access to sentiment labels from validation or test data.

For each domain, candidate terms were filtered using four eligibility criteria: document frequency of at least two, non-zero sentiment score, token length greater than three characters, and a maximum within-domain TF-IDF value greater than 0.35. The common threshold of 0.35 was determined through an elbow-style diagnostic performed exclusively on candidate terms from the sentiment inner-training subset. As shown in Figure 3, the selected value lies within the common transition region across domains, reducing the dense set of low-TF-IDF candidates while retaining sufficient lexical coverage. Neither validation nor held-out test instances or their sentiment labels were used in this diagnostic. The resulting lexicons contained 9 terms for Health, 79 for Media, 78 for Politics, 155 for Society, and 138 for Sports.

Each retained lexicon entry was associated with a sentiment score, denoted as $Score_w$, and a lexical importance weight, denoted as $Weight_w$. The latter combines sentiment strength and TF-IDF-based lexical relevance. During sentiment classification, only the lexicon corresponding to the active domain was used to construct lexical features.

7.3. Topic Detection Experiments

Topic detection experiments evaluated both classical machine learning models and transformer-based architectures. The task involved assigning each tweet to one of the five thematic domains.

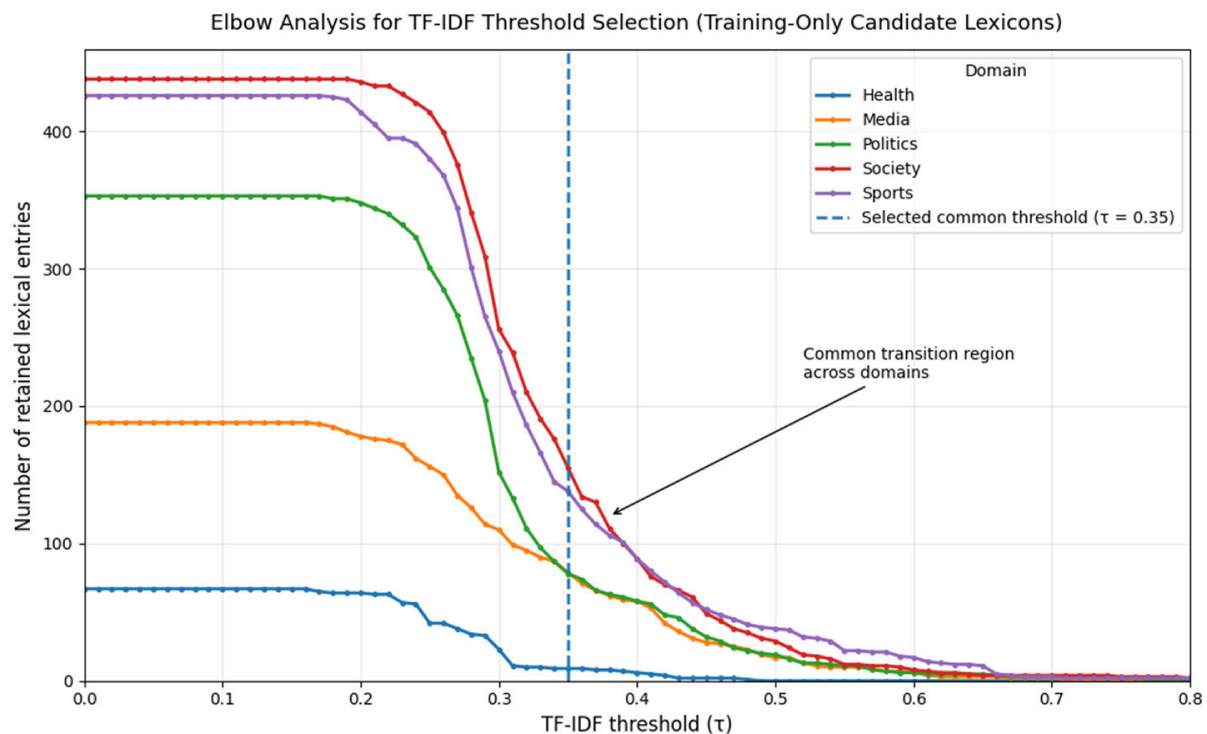


Figure 3. Elbow-style diagnostic for determining the common TF–IDF threshold using candidate terms from the sentiment inner-training subset only.

7.3.1. Experimental Setup for Classical Machine Learning Models

Four classical classifiers were evaluated: Multinomial Naive Bayes, Logistic Regression, Linear Support Vector Machine, and Random Forest. Two feature configurations were examined. In the baseline configuration, tweets were represented using TF–IDF vectors. In the enriched configuration, TF–IDF vectors were augmented with five domain-alignment features, one for each domain-specific lexicon.

The domain-alignment features summarize the lexical compatibility of a tweet with the Health, Media, Politics, Society, and Sports lexicons. For Logistic Regression, Linear SVM, and Random Forest, signed alignment scores were used. Since Multinomial Naive Bayes requires non-negative input values, absolute alignment scores were used in its enriched configuration. These alignment features were computed from the domain-specific lexicons constructed from the sentiment inner-training subset described in Section 7.2; no validation or held-out test tweet contributed to their construction.

Classical configurations were compared on the validation subset using Macro-F1. The held-out test set was used only for final evaluation after the validation-based selection stage.

7.3.2. GreekBERT and Late-Fusion Models

Transformer-based topic detection experiments employed GreekBERT, a pretrained transformer model for Modern Greek. Tweets were tokenized and padded or truncated to a maximum sequence length of 128 tokens.

In the GreekBERT baseline, the contextual representation associated with the [CLS] token was passed to a classification layer for five-class topic prediction. In the late-fusion configuration, the [CLS] representation was concatenated with the five standardized domain-alignment features before classification. The GreekBERT encoder and the late-fusion classification layer were fine-tuned jointly.

For both transformer configurations, checkpoint selection was based on validation loss with early stopping. The late-fusion GreekBERT model was selected as the final topic router because it achieved the strongest validation performance among the evaluated topic-detection configurations. The validation-selected checkpoint was subsequently evaluated once on the held-out topic test set.

7.4. Sentiment Classification Experiments

The sentiment classification experiments evaluated the proposed lexicon-guided structured representation and compared it with stronger baseline approaches. The task was formulated as binary classification between negative and positive sentiment. The evaluated models included TF-IDF combined with a Linear SVM, GreekBERT fine-tuned for binary sentiment classification, a matrix-only CNN using the token-level lexical representation, a global-feature MLP using only aggregate lexical descriptors, and the proposed fusion CNN.

7.4.1. Lexicon-Guided Structured Representation

Each tweet t was represented using two complementary components: $R(t) = \{X_t, f_t\}$. The first component, X_t , is a token-level multi-channel lexicon-guided matrix with dimensions 20×6 . Each row corresponds to a token position, while the six feature channels encode $Score_w$, $Weight_w$, polarity sign, lexicon membership, absolute sentiment score, and the product of sentiment score and lexical weight. The second component, f_t , is a vector of ten global lexical descriptors. These descriptors include score and weight sums and means, positive and negative match counts, total matched-token count, lexicon coverage ratio, maximum positive score, and minimum negative score.

The representation was constructed using the lexicon activated by the corresponding domain. All lexical features were generated using the leakage-free lexicons described in Section 7.2.

7.4.2. Fusion CNN Architecture and Training Procedure

The proposed model consists of a convolutional neural network operating on the token-level matrix X_t . Two one-dimensional convolutional layers extract local patterns from the lexical feature channels. The resulting latent representation is concatenated with the global descriptor vector f_t before the final classification layers. The complete architectural specification and training configuration of the proposed lexicon-guided fusion CNN are summarized in Table 1. The model contains 29,793 trainable parameters, providing a compact neural architecture for the sentiment classification component.

The detailed CNN architecture and training configuration are summarized in Table 1. The model was trained using the validation-based protocol described in Section 7.4.1, where the checkpoint with the lowest validation loss was retained and evaluated once on the held-out sentiment test set.

The matrix-only CNN and global-feature MLP ablations were trained and evaluated using the same fixed data partitions and validation-based model-selection protocol. This design allowed the individual contribution of token-level and aggregate lexical information to be examined.

7.5. Oracle-Topic and Predicted-Topic Evaluation

Two sentiment evaluation settings were considered to distinguish the controlled performance of the sentiment component from the performance of the complete topic-aware pipeline.

Table 1. Architecture and training configuration of the proposed lexicon-guided fusion CNN.

Component	Configuration	Component	Configuration
Input sequence length	20 tokens	Dropout	0.30, applied after the 64-unit CNN dense layer and after the 32-unit dense fusion layer
Token-level feature channels	6	Output layer	1 logit; sigmoid probability threshold of 0.50 for class prediction
Token-level channels	Score_w , Weight_w , polarity sign, lexicon membership, $ \text{Score}_w $, $\text{Score}_w \times \text{Weight}_w$	Loss function	Binary cross-entropy with logits
Global lexical descriptors	10	Optimizer	Adam
Convolutional layer 1	32 filters, kernel size 3, padding 1, ReLU	Learning rate	0.001
Pooling layer 1	Max pooling, kernel size 2	Batch size	32
Convolutional layer 2	64 filters, kernel size 3, padding 1, ReLU	Maximum epochs	30
Pooling layer 2	Max pooling, kernel size 2	Early stopping	Patience = 5 epochs, based on validation loss
Flattened CNN representation	320 units	Checkpoint selection	Lowest validation loss
Dense layer after CNN	64 units, ReLU	Random seed	42
Fusion mechanism	Concatenation of 64-dimensional CNN representation with 10 global lexical descriptors	Reproducibility controls	Fixed seeds; deterministic cuDNN configuration; deterministic PyTorch algorithms
Dense fusion layer	32 units, ReLU	Trainable parameters	29,793
Data partitions	1004 inner-training/112 validation/274 held-out test	Evaluation metrics	Accuracy and Macro-F1

In the reference-domain (oracle-topic) setting, collection-derived domain labels were used to activate the corresponding domain-specific lexicon. This setting evaluates the sentiment representation under the reference domain assignment available in the dataset and provides a controlled reference estimate for the sentiment component. In the predicted-domain setting, the validation-selected GreekBERT late-fusion topic classifier generated domain predictions for the held-out sentiment tweets. These predicted domains were then used to activate the corresponding lexicons. The sentiment fusion CNN was not retrained or tuned in this setting. Instead, the same validation-selected sentiment checkpoint was used only for inference.

This comparison quantifies the extent to which topic-routing errors propagate to the sentiment classification stage and evaluates the proposed framework under a realistic end-to-end configuration.

7.6. Reproducibility

All experiments were conducted using a fixed random seed of 42. The neural models were implemented in PyTorch (version 2.11.0), while classical machine learning models

were implemented using scikit-learn. GreekBERT experiments were implemented using the Hugging Face Transformers library. GPU acceleration was used for the transformer experiments and for the reported CNN training runs.

No held-out test instance contributed to lexicon construction, threshold selection, feature engineering, hyperparameter selection, early stopping, checkpoint selection, or model selection. All reported final results correspond to a single evaluation of validation-selected models on the respective held-out test sets.

8. Results and Discussion

8.1. Topic Detection Results

Topic detection was evaluated using the fixed inner-training, validation, and held-out test partitions described in Section 7. The validation subset was used exclusively for configuration selection and neural checkpoint selection, whereas the held-out test set was used once for final evaluation. Table 2 reports the validation results used for model selection. For the neural models, the reported validation Accuracy and Macro-F1 values are descriptive; checkpoint selection was based on validation loss, as specified in Section 7.3.2.

Table 2. Validation results for topic-detection model selection.

Model/Configuration	Validation Accuracy	Validation Macro-F1
Linear SVM – TF-IDF + signed alignment	0.8839	0.8435
Linear SVM – TF-IDF	0.8839	0.8235
Logistic Regression – TF-IDF	0.8750	0.8174
Random Forest – TF-IDF	0.8571	0.8046
Random Forest – TF-IDF + signed alignment	0.8661	0.8035
Logistic Regression – TF-IDF + signed alignment	0.8393	0.7956
MultinomialNB – TF-IDF	0.8393	0.7381
MultinomialNB – TF-IDF + absolute alignment	0.6429	0.5503
GreekBERT baseline	0.8839	0.8926
GreekBERT late fusion	0.9018	0.8650

Note: Classical configurations were selected using validation Macro-F1. GreekBERT checkpoints were selected using validation loss; therefore, the displayed validation Macro-F1 values were not used as the neural selection criterion. Although the GreekBERT baseline displayed a higher validation Macro-F1 at its selected checkpoint, model selection for both transformer configurations was predefined to use validation loss. The late-fusion checkpoint achieved the lower validation loss and was therefore selected without reference to held-out performance.

Among the classical configurations, the Linear SVM with TF-IDF and signed domain-alignment features achieved the highest validation Macro-F1 (0.8435) and was therefore retained as the best validation classical configuration for comparison. The enriched representation improved the selected SVM configuration relative to TF-IDF alone by 0.0200 Macro-F1, while its validation accuracy remained unchanged. The effect of lexical alignment was not uniform across all classifiers: it improved the selected SVM configuration but reduced validation Macro-F1 for Logistic Regression, Random Forest, and MultinomialNB. This pattern indicates that the utility of the alignment features depends on how each classifier exploits the combined sparse TF-IDF and low-dimensional lexical feature space.

For the transformer models, the GreekBERT baseline checkpoint was selected at epoch 2 with a validation loss of 0.3628, whereas the late-fusion checkpoint was selected at epoch 3 with a lower validation loss of 0.3291. Although a later late-fusion epoch achieved a higher validation Macro-F1, it was not retained because its validation loss was worse. This protocol avoids selecting a neural checkpoint retrospectively on the basis of a secondary validation metric.

Table 3 presents the final held-out topic-test performance. The results show that all models except the enriched MultinomialNB configuration achieved strong performance, with Macro-F1 values above 0.79. The GreekBERT late-fusion router obtained the best held-out performance, reaching an Accuracy of 0.9162 and a Macro-F1 of 0.9116. Moreover, the strongest classical result was obtained by the TF-IDF-only Linear SVM (Accuracy = 0.9031; Macro-F1 = 0.8815). This differs from the validation-based classical selection outcome, where the signed-alignment SVM configuration was selected. The difference is reported transparently and reflects the expected variability associated with a finite validation subset; no configuration was reselected using held-out test performance.

Table 3. Held-out test performance for topic detection.

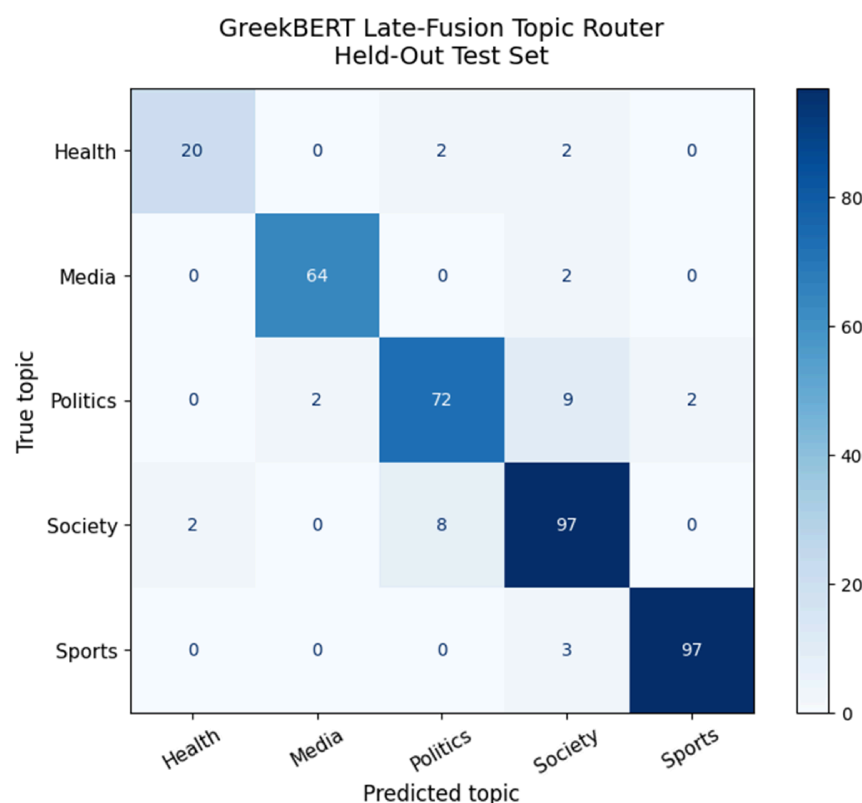
Model/Configuration	Held-Out Accuracy	Held-Out Macro-F1
Linear SVM – TF-IDF + signed alignment	0.8927	0.8727
Linear SVM – TF-IDF	0.9031	0.8815
Logistic Regression – TF-IDF	0.8979	0.8771
Random Forest – TF-IDF	0.8822	0.8631
Random Forest – TF-IDF + signed alignment	0.8848	0.8690
Logistic Regression – TF-IDF + signed alignment	0.8508	0.8311
MultinomialNB – TF-IDF	0.8482	0.7955
MultinomialNB – TF-IDF + absolute alignment	0.6937	0.6231
GreekBERT baseline	0.8743	0.8727
GreekBERT late fusion	0.9162	0.9116

The GreekBERT late-fusion router exceeded the GreekBERT baseline by 0.0419 in Accuracy and 0.0389 in Macro-F1 on the held-out test set. It also exceeded the strongest classical held-out baseline by 0.0131 in Accuracy and 0.0301 in Macro-F1. The effect of lexicon-derived domain-alignment features was therefore model-dependent. In the classical models, the additional features did not produce a consistent held-out improvement over TF-IDF alone, despite improving validation Macro-F1 for the best validation classical SVM configuration. In contrast, the late-fusion GreekBERT configuration achieved a clear held-out gain over the GreekBERT baseline. This pattern suggests that, in the present experimental setting, the low-dimensional lexical alignment signals were more effectively exploited when combined with contextual transformer representations than when appended directly to sparse TF-IDF vectors. Accordingly, the lexicons should not be interpreted as universally improving all topic classifiers; rather, their contribution is most evident in the late-fusion router used for the end-to-end evaluation.

The class-wise results for the selected late-fusion router are summarized in Table 4. Performance was consistently high for Media and Sports, with F1-scores of 0.9697 and 0.9749, respectively. Health, the smallest domain, achieved an F1-score of 0.8696. The lower, though still strong, F1-score for Politics (0.8623) was mainly associated with confusion with Society, while Society achieved an F1-score of 0.8818. The confusion matrix is reported in Figure 4 and shows that the most frequent routing errors occurred between Politics and Society: nine Politics tweets were assigned to Society, while eight Society tweets were assigned to Politics. This pattern is consistent with thematic overlap between political and broader societal discourse in short social media texts.

Table 4. Class-wise held-out performance of the selected GreekBERT late-fusion topic router.

Domain	Precision	Recall	F1-Score	Support
Health	0.9091	0.8333	0.8696	24
Media	0.9697	0.9697	0.9697	66
Politics	0.878	0.8471	0.8623	85
Society	0.8584	0.9065	0.8818	107
Sports	0.9798	0.97	0.9749	100
Macro Average	0.919	0.9053	0.9116	382

**Figure 4.** Confusion matrix of the validation-selected GreekBERT late-fusion topic router on the held-out topic test set. Rows indicate reference domains and columns indicate predicted domains.

8.2. Sentiment Classification Results and Baseline Comparison

This section evaluates the proposed lexicon-guided fusion CNN in a unified multi-domain sentiment-classification setting. A single model was trained across all five domains, while domain awareness was retained through activation of the corresponding domain-specific lexicon for each tweet. The proposed model was compared with two strong sentiment-classification baselines: a TF-IDF representation combined with a Linear SVM and a fine-tuned GreekBERT model. All models were evaluated on the same fixed held-out sentiment test set comprising 274 non-neutral tweets. As described in Section 7, the proposed model was selected using validation loss; the validation-selected checkpoint was obtained at epoch 1 and was evaluated once on the held-out test set. The overall results are presented in Table 5.

The proposed fusion CNN achieved the strongest overall performance, with an Accuracy of 0.8029 and a Macro-F1 of 0.7883. Compared with the TF-IDF + Linear SVM baseline, it improved Accuracy by 0.0292 and Macro-F1 by 0.0294. Compared with GreekBERT, it improved Accuracy by 0.0328 and Macro-F1 by 0.0240. These gains indicate that the explicit lexicon-guided representation provides useful information beyond both sparse lexical TF-IDF features and contextual transformer representations in this dataset.

Table 5. Held-out sentiment-classification performance of the proposed fusion CNN and baseline models.

Model	Held-Out Accuracy	Held-Out Macro F1
TF-IDF + Linear SVM	0.7737	0.7589
GreekBERT	0.7701	0.7643
Proposed fusion CNN	0.8029	0.7883

The result should be interpreted in relation to the task setting. GreekBERT provides a strong general-purpose contextual baseline, whereas the proposed model incorporates domain-conditioned lexical evidence derived from the activated sentiment lexicon. Unlike the TF-IDF baseline, the proposed representation retains token-position information through the multi-channel lexical matrix while also incorporating aggregate lexical descriptors. The fusion architecture can therefore combine localized lexicon-guided patterns with tweet-level evidence that would not be available from aggregate descriptors alone. The improvement of the proposed fusion CNN supports the value of this combined representation in the present dataset. The specific contribution of the two representation components is examined through the ablation study in Section 8.3.

The class-wise held-out results for the validation-selected proposed fusion CNN are reported in Table 6.

Table 6. Class-wise held-out performance of the proposed fusion CNN.

Sentiment Class	Precision	Recall	F1-Score	Support
Negative	0.7807	0.9182	0.8439	159
Positive	0.8506	0.6435	0.7327	115
Macro Average	0.8157	0.7809	0.7883	274

The model achieved strong recall for negative tweets (0.9182), correctly identifying 146 of 159 negative instances. Positive tweets were identified with higher precision (0.8506) but lower recall (0.6435), with 74 of 115 positive instances correctly classified. As shown in Figure 5, 41 positive tweets were assigned to the negative class, whereas 13 negative tweets were assigned to the positive class. Under the fixed decision threshold, the model more frequently assigned positive tweets to the negative class than negative tweets to the positive class. This observed asymmetry is examined further through the qualitative error analysis in Section 8.6.

8.3. Ablation Study

To examine the contribution of the two components of the proposed lexicon-guided representation, an ablation study was conducted using the same fixed inner-training, validation, and held-out test partitions. The full representation combines a token-level multi-channel lexical matrix and a vector of global lexical descriptors. Two reduced configurations were evaluated: (i) a structured matrix-only CNN, which receives only the token-level representation, and (ii) a global descriptors-only MLP, which receives only the aggregate lexical feature vector. All neural models were selected using validation loss and evaluated once on the held-out sentiment test set. All results are shown in Table 7.

The global descriptors-only MLP achieved stronger performance than the structured matrix-only CNN, reaching a Macro-F1 of 0.7761 compared with 0.7526. This result shows that aggregate lexical statistics provide a substantial sentiment signal in the present dataset. However, neither reduced configuration matched the performance of the proposed fusion CNN.

Compared with the structured matrix-only CNN, the proposed fusion model improved Accuracy by 0.0292 and Macro-F1 by 0.0357. Compared with the global descriptors-only MLP, it improved Accuracy by 0.0109 and Macro-F1 by 0.0122. The latter improvement is smaller but consistent across both evaluation metrics.

Overall, the ablation results support the complementary role of the two representation levels. The global descriptors summarize tweet-level lexical evidence, whereas the token-level multi-channel matrix preserves the position and local arrangement of lexicon-derived signals. Their fusion enables the model to combine aggregate sentiment characteristics with localized lexical patterns, yielding the strongest held-out performance.

Table 7. Ablation results for the lexicon-guided sentiment representation.

Model/Input Configuration	Selected Epoch	Best Validation Loss	Held-Out Accuracy	Held-Out Macro-F1
Structured matrix-only CNN	1	0.5173	0.7737	0.7526
Global descriptors-only MLP	17	0.4884	0.7920	0.7761
Proposed fusion CNN	1	0.4957	0.8029	0.7883

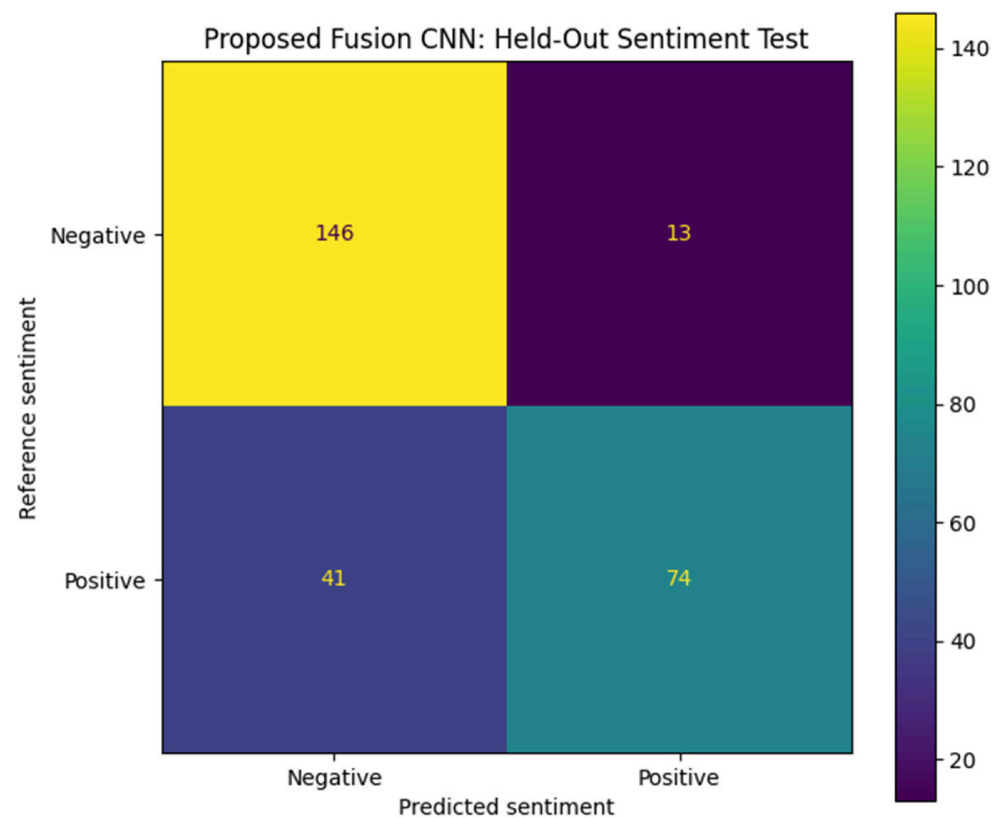


Figure 5. Confusion matrix of the validation-selected proposed fusion CNN on the held-out sentiment test set. Rows indicate reference sentiment labels and columns indicate predicted labels.

8.4. Domain-Wise Sentiment Results

To examine the behavior of the proposed fusion CNN across thematic domains, held-out sentiment performance was also calculated separately for each domain. Table 8 reports the number of test tweets, class composition, Accuracy, and Macro-F1. These results are descriptive because the domain-specific test subsets are small and, in several cases, strongly imbalanced.

Table 8. Domain-wise held-out sentiment performance of the proposed fusion CNN.

Domain	Test Tweets	Negative Tweets	Positive Tweets	Accuracy	Macro-F1
Health	15	14	1	0.9333	0.4828
Media	37	4	33	0.7297	0.5027
Politics	64	64	0	1	—
Society	85	52	33	0.6235	0.5355
Sports	73	25	48	0.8356	0.8262

The domain-level results should be interpreted together with the class distributions. The Politics subset contains only negative tweets; consequently, its perfect Accuracy does not assess discrimination between positive and negative sentiment, and Macro-F1 is not reported. Similarly, Health contains only one positive instance and Media only four negative instances. Their Accuracy values are therefore sensitive to a small number of errors, while Macro-F1 provides a more informative but still unstable estimate.

Among the domains with both sentiment classes represented in more substantial numbers, Sports achieved the strongest balanced result (Macro-F1 = 0.8262). Society was more challenging (Macro-F1 = 0.5355), suggesting that the available representation captured its held-out sentiment patterns less consistently. This interpretation is tentative because the per-domain subsets remain limited. The domain-wise results therefore complement, rather than replace, the aggregate held-out evaluation reported in Section 8.2.

8.5. Oracle-Topic and Predicted-Topic Results

The results reported in Sections 8.2–8.4 use reference-domain labels derived from the collection query/hashtag to activate the corresponding domain-specific lexicon. This oracle-topic setting isolates the performance of the lexicon-guided sentiment representation when the corresponding domain-specific knowledge resource is selected. To evaluate the complete topic-aware pipeline, the proposed fusion CNN was also evaluated using domains predicted by the validation-selected GreekBERT late-fusion topic router. The sentiment fusion CNN was not retrained or retuned for this setting; the same validation-selected checkpoint used in the oracle-topic evaluation was used only for inference.

The predicted-topic setting reduced Accuracy by 0.0292 and Macro-F1 by 0.0314 relative to oracle-topic activation, as shown in Table 9. This decrease quantifies the effect of replacing reference-domain lexicon activation with automatically predicted-domain activation. Nevertheless, the end-to-end pipeline retained an Accuracy of 0.7737 and a Macro-F1 of 0.7569, indicating that the sentiment representation remains operational when domain-specific lexicons are selected automatically rather than from reference-domain labels.

Table 9. Held-out sentiment performance under oracle-topic and predicted-topic lexicon activation.

Sentiment Setting	Held-Out Accuracy	Held-Out Macro-F1
Oracle-topic lexicon activation	0.8029	0.7883
Predicted-Topic Lexicon Activation	0.7737	0.7569
Difference	−0.0292	−0.0314

Because the active domain-specific lexicon is used to construct both token-level and global lexical features, a different domain prediction can alter the representation supplied to the sentiment classifier. The oracle-topic result should therefore be interpreted as a controlled estimate of the sentiment component under reference-domain activation, whereas the predicted-topic result represents the realistic performance of the full framework.

8.6. Lexical Analysis and Qualitative Error Analysis

8.6.1. Analysis of Domain-Specific Lexicons

To support the qualitative error analysis presented in the following subsection, this section first examines the characteristics of the leakage-free domain-specific lexicons used by the proposed framework. Table 10 summarizes the lexicons constructed exclusively from the sentiment inner-training subset, reporting their size, polarity distribution, and average sentiment statistics. These characteristics provide the contextual basis for interpreting how lexical coverage, polarity balance, and domain-specific sentiment associations influence both successful predictions and residual classification errors. The reported statistics describe the lexical resources used in the experiments and should not be interpreted as general-purpose sentiment dictionaries.

Table 10. Summary of leakage-free domain-specific sentiment lexicons.

Domain	Lexicon Size	Positive-Score Terms	Negative-Score Terms	Mean $ Score_w $	Mean $ Weight_w $
Health	9	0	9	2.7623	1.1715
Media	79	71	8	2.1916	0.9976
Politics	78	0	78	2.7578	1.2648
Society	155	52	103	1.781	0.7755
Sports	138	79	59	2.5467	1.1804

The lexicons vary substantially in both size and polarity composition. Health produced the smallest lexicon, with only nine retained terms after applying the leakage-free eligibility criteria. Politics contained exclusively negative-score terms, whereas Media was predominantly positive. Society and Sports contained both positive and negative entries, although Society was skewed toward negative terms and Sports was comparatively more balanced.

This variation reflects the composition of the sentiment-annotated inner-training tweets from which the lexical scores were estimated. In particular, a positive or negative $Score_w$ represents the average annotation of tweets containing a term within its domain-specific training corpus; it does not imply that the term has a fixed polarity across all Greek social media contexts. This distinction is important for entity-linked terms and general lexical items whose sentiment contribution depends on the surrounding discourse.

Selected entries illustrate the domain-conditioned nature of the lexical resources. For example, the Sports lexicon includes positive terms such as *θεαρά* ($Score_w = 5.0000$; $Weight_w = 3.2546$), *συγχαρητηρια* (4.3000; 2.0621), and *αθλητης* (3.7778; 2.2723). In Politics, entries such as *κυβερνηση* (−4.8333; 2.1695) and *τσιπρας* (−4.2500; 2.2015) received negative scores in the observed training discourse. The Health lexicon includes *εμβολιασμενοι* (−2.7500; 1.0999), reflecting the sentiment composition of vaccination-related tweets in the corresponding subset. These examples demonstrate that the lexicons encode domain-conditioned sentiment associations rather than generic word-level polarity.

The small Health lexicon and the strongly skewed polarity distributions in some domains also indicate a limitation of the approach. Under strict leakage-free construction and frequency-based filtering, lexical coverage depends on the amount and composition of the available domain-specific training evidence. This limitation is consistent with the cautious interpretation of the domain-wise results in Section 8.4 and motivates the qualitative analysis of residual errors in the following subsection.

8.6.2. Qualitative Error Analysis

A targeted qualitative inspection of representative held-out misclassifications was conducted to identify recurrent challenges for the lexicon-guided representation. The analysis in Table 11 is illustrative rather than exhaustive. It focuses on cases in which lexical evidence is incomplete, context-dependent, or potentially misleading. Tweet excerpts were normalized and shortened where necessary. The tweet excerpts are retained in Greek, with usernames, URLs, and non-essential identifying elements removed or normalized. They are not translated in the table because translation could alter lexical form, wordplay, entity references, and pragmatic cues that are central to the present error analysis.

Table 11. Illustrative held-out sentiment-classification errors.

Domain	Normalized Tweet Excerpt	Reference Label	Predicted Label	Main Challenge	Interpretation
Society	“δεν ειστε εργαζομενοι εσεις σας νοιαζει...”	Negative	Positive	Context-dependent stance	The tweet expresses a critical stance, but its polarity is conveyed through the broader proposition rather than through a small set of strongly negative lexical cues.
Society	“αλληλεγγυη στον αγωνα των απεργων...”	Negative	Positive	Potentially misleading positive lexical cue	The term <i>αλληλεγγυη</i> may contribute positive lexical evidence, although the overall annotation is negative. This illustrates that a locally positive term does not necessarily determine tweet-level sentiment.
Society	“μπραβο κυριακο μπραβο λεβεντη μου”	Positive	Negative	Entity-dependent and pragmatically ambiguous evaluation	The sentiment interpretation depends on the target entity and pragmatic context. Surface praise alone may be insufficient to determine whether the expression is literal or contextually marked.
Sports	“φερναμε καλαθη εστω και τωρα χωρις αλλη μεταγραφη...”	Positive	Negative	Implicit positive evaluation	The positive reference label appears to rely on approval of a desired sporting outcome. The tweet expresses satisfaction through event-specific context rather than through explicit high-weight positive terms.
Media	“ωραιος ηθοποιος χαραλαμποπουλος...”	Positive	Negative	Entity-dependent evaluation and lexical coverage	The tweet expresses positive evaluation of a named actor, but the sentiment is tied to the entity and a short evaluative phrase. Sparse or absent matching lexical evidence can leave the representation dominated by other cues.

The examined errors show that the structured lexical representation is most informative when sentiment is expressed through lexical cues that are represented in the activated domain-specific lexicon. In contrast, errors arise when sentiment depends on pragmatic interpretation, an entity-specific stance, or information distributed across the full proposition rather than encoded in individual lexical items.

The examples also clarify the interpretability boundary of the proposed approach. The lexical matrix makes it possible to inspect which domain-specific lexical signals are available to the model, but it does not resolve all forms of contextual meaning. In particular, short social media texts may express sentiment implicitly, through irony, sarcasm, or other pragmatic framing, or through references to people and events whose polarity is not stable

across contexts. These cases motivate future extensions that combine the explicit lexical representation with richer contextual and entity-aware modeling.

9. Contribution

This study contributes a lexicon-guided structured semantic representation for sentiment classification in Greek social media. Rather than reducing domain-specific lexical evidence to aggregate document-level statistics alone, the proposed representation encodes token-level lexical attributes in a multi-channel matrix and combines them with global lexical descriptors. This design preserves the position and local arrangement of explicit lexical signals while retaining tweet-level information about lexical coverage, polarity distribution, and sentiment intensity.

A second contribution is the topic-aware lexicon activation mechanism. Domain routing selects the corresponding domain-specific lexicon before the sentiment representation is constructed, enabling lexical evidence to be interpreted within the thematic context of the tweet. The framework should therefore be understood as a lexicon-guided structured feature representation, rather than as a knowledge-graph or graph-reasoning approach.

Methodologically, the study introduces a leakage-free evaluation protocol in which domain-specific sentiment lexicons are reconstructed exclusively from the inner-training subset. The TF-IDF threshold and lexicon-construction criteria are determined exclusively from the inner-training subset, while validation data are used for model selection and checkpoint selection. This protocol addresses the risk that sentiment labels from evaluation instances could otherwise influence lexical scores, lexical weights, or feature construction.

The empirical evaluation compares the proposed fusion CNN with TF-IDF + Linear SVM and fine-tuned GreekBERT baselines under the same held-out sentiment test protocol. In the present dataset, the fusion CNN achieved the strongest held-out result, with an Accuracy of 0.8029 and a Macro-F1 of 0.7883. The ablation analysis further showed that both representation components contribute complementary information: the global lexical descriptors provided a strong aggregate signal, while their fusion with the token-level lexical matrix produced the best overall performance.

Finally, the study distinguishes between controlled reference-domain activation and predicted-domain end-to-end activation. Under reference-domain activation, the proposed fusion CNN achieved a Macro-F1 of 0.7883, whereas activation through predicted domains produced a Macro-F1 of 0.7569. This comparison makes the effect of topic-routing errors explicit and separates the performance of the sentiment representation itself from the performance of the complete topic-aware pipeline.

Overall, the findings show that explicit domain-specific lexical knowledge can be incorporated into a compact neural sentiment-classification architecture through an interpretable structured representation. Within the present Greek social media dataset, the approach demonstrates that explicit domain-specific lexical knowledge can complement both sparse TF-IDF features and contextual transformer representations while making the contribution and limitations of the lexical evidence available for direct inspection.

10. Limitations

This study is based on a single manually annotated Greek social media dataset collected during a specific time period and organized into five reference domains derived from the collection query/hashtag. Although the dataset supports controlled comparison across the proposed configurations, the findings should not be assumed to generalize directly to other Greek social media periods, platforms, topic taxonomies, or languages. External validation on additional public datasets and temporally distinct collections is therefore required.

The domain-specific lexicons are corpus-derived resources rather than general-purpose sentiment dictionaries. Their lexical scores reflect the average annotation of the inner-training tweets in which a term occurred, and may not remain stable across different discourse contexts. Lexicon size and polarity composition also varied substantially across domains. In particular, the strict leakage-free eligibility criteria yielded a small Health lexicon, while some domains exhibited strongly skewed positive or negative lexical-score distributions. These properties can limit lexical coverage and may affect the representation differently across domains.

The primary sentiment comparison was conducted under reference-domain activation, which was used as a controlled setting to isolate the effectiveness of the proposed lexicon-guided sentiment representation when the relevant domain-specific knowledge resource is selected correctly. The study additionally reports a predicted-domain evaluation for the proposed fusion CNN to quantify the performance of the complete topic-aware pipeline. However, equivalent predicted-domain evaluations were not conducted for the TF-IDF + Linear SVM and GreekBERT sentiment baselines. The present study therefore does not establish whether the relative performance differences observed under reference-domain activation are preserved to the same extent when all systems operate under automatically predicted domain assignments.

The qualitative error analysis further indicates that the current representation does not fully resolve pragmatic ambiguity, entity-dependent evaluation, implicit sentiment, or context-sensitive lexical meaning. Short social media texts may express polarity through proposition-level meaning, references to people or events, or discourse context that is not captured by the available lexical signals. The model is therefore best interpreted as an explicit lexical knowledge-guided component rather than as a complete solution to contextual sentiment understanding.

Finally, although the proposed fusion CNN achieved the strongest held-out sentiment result among the evaluated models in the present experimental setting, the study does not provide a comprehensive comparison of computational cost, inference latency, or hardware efficiency against transformer baselines. The parameter analysis reported in the experimental section provides an architecture-level indication of model size, but it should not be interpreted as a full efficiency benchmark.

11. Conclusions and Future Work

This study presented a topic-aware, lexicon-guided structured representation for sentiment classification in Greek social media texts. The framework activates a domain-specific sentiment lexicon and transforms its lexical evidence into a token-level multi-channel matrix, which is combined with global lexical descriptors in a fusion CNN. This design preserves explicit lexical signals at both local and tweet-level scales while making the lexical evidence used by the model available for inspection.

The evaluation followed a leakage-free protocol in which the domain-specific lexicons were constructed exclusively from the sentiment inner-training subset, while validation data were used for model selection and the held-out test set was reserved for final evaluation. Under reference-domain activation, the proposed fusion CNN achieved the strongest held-out sentiment performance among the evaluated models, reaching an Accuracy of 0.8029 and a Macro-F1 of 0.7883. The ablation study showed that the token-level matrix and global lexical descriptors provide complementary information, with their fusion producing the best overall result in the present experimental setting.

The topic-detection evaluation further showed that GreekBERT late fusion was the strongest final router, achieving a held-out Accuracy of 0.9162 and a Macro-F1 of 0.9116. When the sentiment pipeline used predicted-domain rather than reference-domain lexi-

con activation, the proposed framework retained a Macro-F1 of 0.7569. This end-to-end result distinguishes the controlled evaluation of the sentiment representation from the performance of the complete topic-aware pipeline and makes the effect of domain-routing errors explicit.

The lexical and qualitative analyses indicate both the value and the boundaries of the approach. Domain-specific lexicons provide interpretable evidence for sentiment classification, but their coverage and polarity associations remain dependent on the available training corpus. Errors were particularly associated with pragmatic ambiguity, entity-dependent evaluation, implicit sentiment, and context-sensitive lexical meaning, which are difficult to capture through lexical evidence alone.

Future work should evaluate the framework on additional public and temporally distinct Greek datasets, and extend the comparison so that all sentiment baselines are also assessed under predicted-domain activation. Further directions include adaptive lexicon expansion, automated updating of domain-specific lexical resources, and the integration of the explicit lexical representation with entity-aware and contextual modeling mechanisms. These extensions could improve resilience to evolving vocabulary, pragmatic language use, and topic-routing uncertainty while preserving the transparency of the proposed lexical component.

Author Contributions: Conceptualization, K.S. and I.K.S.; methodology, K.S.; software, K.S.; validation, K.S. and I.K.S.; data curation, K.S.; writing—original draft preparation, K.S.; writing—review and editing, I.K.S.; visualization, K.S.; supervision, I.K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated and analyzed during the current study are not publicly available, as they form part of an ongoing doctoral research project. However, the data are available from the corresponding author upon reasonable request for the purposes of academic review, validation, or replication of the reported results.

Acknowledgments: The authors used OpenAI ChatGPT (GPT-5.5 Instant) as an AI-assisted language tool during the preparation of the manuscript for translation support, editorial refinement, linguistic improvement, and structural polishing of selected text passages. All scientific content, experimental design, methodology, results, figures, and interpretations were developed, verified, and reviewed by the authors. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
TF	Term Frequency
IDF	Inverse Document Frequency
SVM	Support Vector Machine
CNN	Convolutional Neural Network
Macro-F1	Macro-averaged F1 score

References

1. Liu, B. *Sentiment Analysis and Opinion Mining*; Morgan & Claypool: San Rafael, CA, USA, 2012.

2. AGiachanou; Crestani, F. Like it or not: A survey of Twitter sentiment analysis methods. *ACM Comput. Surv.* **2016**, *49*, 1–41.
3. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [[CrossRef](#)]
4. Kim, Y. Convolutional neural networks for sentence classification. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014; pp. 1746–1751.
5. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
6. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the NeurIPS, Virtual, 6–12 December 2020; pp. 1877–1901.
7. Peng, X.; Papadopoulos, T.; Soufleri, E.; Giannouris, P.; Xiang, R.; Wang, Y.; Qian, L.; Huang, J.; Xie, Q.; Ananiadou, S. Plutus: Benchmarking large language models in low-resource Greek finance. In Proceedings of the EMNLP, Suzhou, China, 4–9 November 2025; pp. 30164–30190. [[CrossRef](#)]
8. Mastrokostas, C.; Giarelis, N.; Karacapilidis, N. Evaluating monolingual and multilingual large language models for Greek question answering: The DemosQA benchmark. *arXiv* **2026**, arXiv:2602.16811.
9. Zhao, L.; Huang, K.; Ma, X.; Zhu, H.; Yuan, C.; Su, Y. SAGU-Net: Gate-level lexicon–neural fusion via sentiment-aware gated units for social media sentiment analysis. *Appl. Sci.* **2026**, *16*, 4994. [[CrossRef](#)]
10. Lin, Z.; Wu, D.; Yang, X.; Li, L.; Qin, Z. FEAM: A dynamic prompting framework for sentiment analysis with hierarchical convolutional attention. *Front. Phys.* **2025**, *13*, 1674949. [[CrossRef](#)]
11. Hu, M.; Liu, B. Mining and summarizing customer reviews. In Proceedings of the KDD, Seattle, WA, USA, 22–25 August 2004; pp. 168–177.
12. Hamilton, W.L.; Leskovec, J.; Jurafsky, D. Diachronic word embeddings reveal statistical laws of semantic change. In Proceedings of the ACL, Berlin, Germany, 7–12 August 2016; pp. 1489–1501.
13. Pak, A.; Paroubek, P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 17–23 May 2010; European Language Resources Association (ELRA): Paris, France, 2010.
14. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Proceedings of the NeurIPS, Montreal, QC, Canada, 7–12 December 2015; pp. 649–657.
15. Vo, D.T.; Zhang, Y. Target-dependent Twitter sentiment classification with rich automatic features. In Proceedings of the IJCAI, Buenos Aires, Argentina, 25–31 July 2015; pp. 1347–1353.
16. Tang, D.; Qin, B.; Liu, T. Learning semantic representations of users and products for document-level sentiment classification. In Proceedings of the ACL, Beijing, China, 27–29 July 2015; pp. 1014–1023.
17. Sun, C.; Huang, L.; Qiu, X. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 380–385.
18. Koutsikakis, G.; Chalkidis, A.; Androutsopoulos, I.; Malakasiotis, P. GreekBERT: The Greeks visiting Sesame Street. In Proceedings of the SETN, Athens, Greece, 2–4 September 2020.
19. Samaras, L.; García-Barriocanal, E.; Sicilia, M.-A. Sentiment analysis of COVID-19 cases in Greece using Twitter data. *Expert Syst. Appl.* **2023**, *230*, 120577. [[CrossRef](#)] [[PubMed](#)]
20. Mastrokostas, C.; Giarelis, N.; Karacapilidis, N. Social media topic classification on Greek Reddit. *Information* **2024**, *15*, 521. [[CrossRef](#)]
21. Kouloumpis, E.; Wilson, T.; Moore, J. Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of the ICWSM, Barcelona, Spain, 17–21 July 2011.
22. Skoularikis, K.; Savvas, I.; Garani, G.; Kakarontzas, G. Hybrid sentiment analysis approach with domain-specific lexicons with active learning on Greek social media texts. *Soc. Netw. Anal. Min.* **2026**, *16*, 35. [[CrossRef](#)]
23. Kowsari, K.; Heidarysafa, M.; Brown, D.E.; Meimandi, K.J.; Barnes, L.E. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [[CrossRef](#)]
24. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning-based text classification: A comprehensive review. *ACM Comput. Surv.* **2021**, *54*, 1–40. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.