Enhanced Sentiment Analysis Using Extended Corpus and Co-Occurrence Graph

Nirach Romyen [1,*], Herwig Unger [2], and Maleerat Maliyaem [1,*]

[1] Department of Information Technology, Faculty of Information Technology and Digital Innovation, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
[2] Department of Mathematics and Computer Science, University of Hagen, Hagen, Germany
Email: nirach@sci.tu.ac.th (N.R.); herwig.unger@gmail.com (H.U.); maleerat.m@itd.kmutnb.ac.th (M.M.)
*Corresponding author

Abstract*—Lexicon-based sentiment analysis often faces challenges due to the limited coverage and adaptability of existing sentiment lexicons, particularly in domain-specific or resource-scarce contexts. Constructing new lexicons from scratch is labor-intensive and rarely feasible for large-scale applications. To address this gap, we introduce the Enhanced Sentiment Analysis using Extended Corpus and Co-Occurrence Graph (ECG) framework, which leverages co-occurrence graph construction, triadic closure theory, and PageRank-inspired weighting to propagate sentiment values from a small seed set to a broader vocabulary. Our study focuses on nouns extracted from the Harry Potter corpus, which are typically underrepresented in standard sentiment resources. Experimental results demonstrate that ECG achieves effective sentiment propagation within 1500 iterations, reaching stable convergence. Evaluation using the Coefficient of Variation (COV) indicates improved sentiment differentiation, with broader coverage compared to benchmark lexicons such as SentiWordNet, Valence Aware Dictionary and Sentiment Reasoner (VADER), and National Research Council (NRC). For instance, ECG consistently assigned interpretable sentiment values to narrative-specific entities (e.g., "Slytherin", "Voldemort"), which were often missed by conventional lexicons. These findings highlight ECG's scalability, interpretability, and suitability for low-resource scenarios. Beyond literary analysis, the framework can be extended to other domains such as social media, customer feedback, and specialized corpora. Overall, ECG represents a resource-efficient alternative that bridges the limitations of static lexicons and enhances sentiment inference in diverse contexts.

***Keywords*—sentiment analysis, co-occurrence graph, sentiment propagation, triadic closure, lexical expansion**

I. Introduction

Sentiment analysis is a core task in Natural Language Processing (NLP) and Artificial Intelligence (AI), widely applied in areas such as social media monitoring, customer feedback analysis, and opinion mining [1, 2]. Traditional approaches can be broadly divided into two paradigms: machine learning–based models and lexicon-based methods [3]. Machine learning approaches, particularly supervised deep learning models, often achieve high predictive accuracy but require large-scale annotated datasets, which are costly and domain-dependent. They also tend to operate as black boxes, limiting interpretability [4, 5].

Lexicon-based approaches, on the other hand, are valued for their transparency and ease of use, since they rely on predefined dictionaries mapping words to sentiment scores. However, they face a persistent limitation: sentiment lexicons are often incomplete or unavailable in niche or resource-constrained domains. For instance, common lexicons such as SentiWordNet, Valence Aware Dictionary and Sentiment Reasoner (VADER), and National Research Council (NRC) Emotion Lexicon provide broad coverage in general contexts, but fail to capture domain-specific or narrative-driven terms. Building new lexicons manually is both time-consuming and resource-intensive. A comprehensive review of existing lexicon-building methods reveals that most struggle with coverage gaps, domain adaptability, or lack of fine-grained sentiment differentiation.

To address these limitations, recent studies have explored graph-based lexicon expansion, where words are represented as nodes and co-occurrences form edges [6]. Using structural properties such as triadic closure and centrality, sentiment can be propagated from a small seed set of labeled words to unlabeled terms [6, 7]. This approach enables automatic lexicon growth while preserving semantic relationships. Building on this foundation, the present study introduces the Enhanced Sentiment Analysis using Extended Corpus and Co-occurrence Graph (ECG) framework. The method integrates:

- Co-occurrence graph construction to capture narrative-specific word relations.
- Triadic closure principles to strengthen propagation through meaningful local structures [8].
- A PageRank-inspired weighting scheme to emphasize central and influential nodes in sentiment diffusion [9].

Unlike prior work, ECG focuses initially on noun terms, which are typically underrepresented in sentiment

Manuscript received June 18, 2025; revised September 1, 2025; accepted November 9, 2025; published February 10, 2026.

resources but play an essential role in narrative texts. Using only 10–20 seed words, ECG propagates sentiment values iteratively across the graph until convergence.

To evaluate effectiveness, this study applied ECG to the Harry Potter corpus (~1.1M tokens). Results were benchmarked against SentiWordNet, VADER, and NRC across 1500 propagation iterations. The evaluation highlights three major findings:

- Improved coverage—ECG assigns interpretable sentiment values to narrative-specific terms that traditional lexicons miss.
- Enhanced granularity—measured by the Coefficient of Variation (COV), ECG achieves clearer sentiment differentiation, particularly when seed size expands from 10 to 20.
- Baseline superiority—as shown in Table I, ECG consistently produces stronger and more distinct polarity values compared to existing lexicons.

TABLE I. COMPARISON OF COEFFICIENT OF VARIATION (COV) WITH DIFFERENT NUMBERS OF SEED WORDS (AT ITERATION $T = 1500$)

Seed Configuration	Positive Only (%)	Negative Only (%)	Positive & Negative (%)
10 seed words	30.28	31.61	41.17
20 seed words	41.55	41.68	55.75
30 seed words	38.92	39.11	52.46

Beyond narrative texts, the framework demonstrates potential for generalization to other domains (e.g., legal, medical, or technical corpora) and extensibility to additional word classes such as verbs and adjectives. Furthermore, integrating ECG with transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) may enhance contextual sensitivity, offering a hybrid pathway between interpretable lexicon-based analysis and powerful deep learning models.

In summary, ECG bridges the gap between resource-efficient lexicon construction and data-driven adaptability. It contributes a scalable, interpretable, and domain-adaptable method that addresses the limitations of both existing lexicons and opaque machine learning systems.

II. LITERATURE REVIEW

This section reviews theoretical foundations and prior work relevant to ECG, highlighting the limitations of existing sentiment lexicons, survey of lexicon expansion approaches, and the supporting graph-theoretical concepts underpinning our method.

A. Limitations of Existing Lexicons

Lexicon-based sentiment analysis has benefited from widely used resources such as SentiWordNet [10], VADER [11], and the NRC Emotion Lexicon [12]. Despite their popularity, these lexicons face several critical limitations:

- Domain dependency—Coverage and polarity of existing lexicons often fail to transfer to specialized domains such as law, medicine, or narrative fiction.
- Context insensitivity—Fixed polarity scores cannot adapt when sentiment changes with context, register, or genre.
- Multilingual and low-resource constraints—Most lexicons are developed for English and are not easily scalable to languages such as Thai or to corpora with limited annotated resources.
- Static polarity assignment—Sarcasm, irony, or polysemy are rarely addressed, and words are typically assigned a single polarity value regardless of usage.

Our comparative experiments confirm these limitations. As shown in Table I, ECG assigns more distinctive polarity values to narrative-specific terms (e.g., Harry, Ron, Snape) than standard lexicons, which tend to yield weaker or near-neutral scores. This evidence underscores the need for corpus-driven approaches that adapt directly to the target domain.

B. Lexicon Expansion Approaches

Prior work has sought to automate lexicon construction through bootstrapping from seed lists, label propagation over word graphs, and semantic similarity from distributional representations [13, 14]. These methods improve coverage but share critical drawbacks, most still rely on large existing lexicons as initialization, which introduces inherited bias and retains domain dependence.

Recent neural and transformer-based models (e.g., BERT embeddings for sentiment induction [15]) provide richer contextual signals but often operate as black boxes, lacking transparency and interpretability. In contrast, ECG leverages graph-theoretical principles with explicit propagation dynamics, offering both quantitative improvements and transparent mechanisms. By building from a corpus-specific seed rather than from external lexicons, ECG reduces inherited bias and adapts naturally to domain vocabulary.

C. Transformer-Based Sentiment Models.

Recent research has demonstrated the superior performance of transformer architectures for sentiment classification. Supal *et al.* [16] conducted an extensive evaluation of transformer-based models such as BERT, Robustly Optimized BERT Pretraining Approach (RoBERTa), Distilled BERT (DistilBERT), and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), reporting that ELECTRA achieved the highest F1-score and accuracy across diverse datasets, thereby confirming the ability of transformers to capture contextual dependencies and sentiment cues. In the Thai NLP domain, Porkaew *et al.* [17] introduced HoogBERTa, a RoBERTa-based pretrained model employing multi-task learning for Thai sequence labeling tasks such as part-of-speech tagging and named entity recognition. The model achieved strong F1 performance across tasks and highlighted the potential of transformer-based architectures for Thai sentiment understanding. However, while these models demonstrate high accuracy, they depend on large annotated datasets and substantial computational resources. In contrast, the proposed ECG framework complements these models by propagating sentiment through co-occurrence structures, enabling effective sentiment analysis under low-resource conditions.

D. Co-Occurrence Graph

A co-occurrence graph represents words as nodes and connects them with edges when they appear in the same sentence or local context [18]. Unlike simple frequency counts, this structure captures relational semantics: how terms co-occur in meaningful contexts.

Formally, the graph is defined as $G = (V, E)$ where V is the set of nodes (noun terms in this study) and E is the set of edges, weighted by the Dice coefficient to normalize co-occurrence strength between 0 and 1 [19]. To reduce semantic noise, weak associations below a threshold are pruned. This ensures that the resulting network reflects meaningful lexical relationships relevant to sentiment propagation.

The co-occurrence graph serves as the structural backbone of ECG. As later validated in Section IV, it enables polarity values to diffuse effectively, yielding broader sentiment coverage than static lexicons (see Fig. 1).

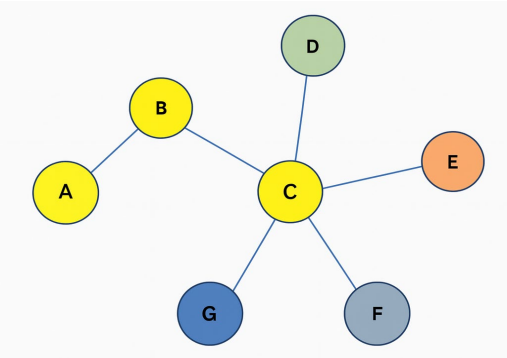

Fig. 1. Co-occurrence graph.

In this research, the co-occurrence graph forms the basis for capturing semantic associations between terms and plays a central role in supporting sentiment propagation throughout the lexical network.

E. Triadic Closure Theory

The principle of Triadic Closure reflects the natural tendency in networks: if two nodes (e.g., B and C) share a strong connection with a common neighbor (A), it is highly likely that a direct link between them will eventually form. This mechanism captures the intuition that "a friend of my friend is also likely to become my friend" [20].

As illustrated in Fig. 2, the left-hand side shows a graph where nodes B and C are not directly connected but both are strongly linked to A. According to Triadic Closure, this structural pattern implies a missing but probable link between B and C. The right-hand side of the figure demonstrates the completed closure, where an additional edge (in red) connects B and C.

Within the ECG framework, this principle becomes crucial for sentiment propagation. Even if two terms do not co-occur directly in the text, their mutual relationship to a shared neighbor allows sentiment cues to pass indirectly between them. This supports:

- Mitigation of data sparsity—inferring connections where direct co-occurrence is rare.
- Improved coverage—enabling unlabeled terms to inherit sentiment values from structurally related neighbors.
- Semantic consistency—ensuring that related terms reinforce one another's polarity orientation.

Beyond the immediate narrative corpus, the use of Triadic Closure also contributes to the generalizability of the ECG method. Many domains—such as legal, medical, and social media texts—are characterized by sparse or specialized vocabularies where co-occurrence information is incomplete. By leveraging Triadic Closure, ECG can bridge gaps in these low-resource settings and infer sentiment relations even when annotated lexicons are unavailable. In addition, because the mechanism relies only on graph structure rather than language-specific resources, it is adaptable across multilingual corpora. This allows ECG to extend beyond English fiction and address domain-specific challenges, such as interpreting sentiment in clinical notes, legal judgments, or product reviews.

Thus, the incorporation of Triadic Closure not only strengthens sentiment propagation within the graph but also enhances the robustness and domain transferability of the framework, demonstrating its applicability beyond the initial narrative testbed.

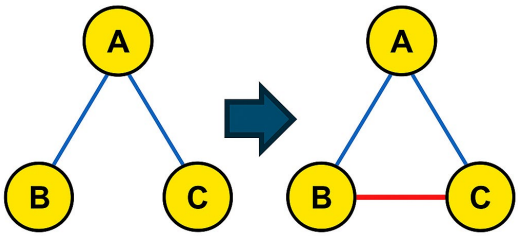

Fig. 2. Triadic closure.

F. PageRank Algorithm

PageRank, originally proposed for ranking web pages [21], evaluates the relative importance of nodes based on their connectivity to other influential nodes. Within the ECG framework, PageRank is applied to the co-occurrence graph so that highly connected terms exert greater influence in sentiment propagation. This ensures that sentiment values spread more effectively through central nodes rather than being dominated by sparsely connected or peripheral terms.

The standard PageRank formulation is given by Eq. (1):

$$PR_i = (1 - \eta) + \eta \sum_{j \in S} \frac{PR_j}{|N_j|} \quad (1)$$

where η is the damping factor (commonly set to 0.85), S is the set of nodes linking to the target node i, and $|N_j|$ is the out-degree of node j. PR_i denotes the PageRank score of node i, and PR_j denotes the PageRank score of each neighboring node j that contributes to i through incoming links.

In practice, incorporating PageRank weighting provides several advantages relevant to the robustness and scalability of the ECG framework:

- Seed size and stopping criteria: By prioritizing highly connected nodes, the propagation stabilizes quickly, reducing sensitivity to the number of initial seeds or arbitrary stopping thresholds. Experimental results show that polarity distributions converge after approximately 1500 iterations, indicating robustness even with a small initial seed set.
- Quantitative evaluation: PageRank-based propagation yields polarity distributions that are more consistent and comparable with existing lexicons (e.g., SentiWordNet, VADER, NRC). This ensures that the quantitative comparison remains meaningful, as the ECG outputs are less influenced by initialization noise.
- Generalizability and scalability: PageRank is designed to handle large-scale networks efficiently, supporting the extension of the ECG framework to broader domains such as legal and medical corpora, as well as to multilingual settings. The graph-based weighting scheme ensures that sentiment propagation remains stable even when vocabulary size and co-occurrence relations increase significantly.

As shown in Fig. 3, the application of PageRank leads to a consistent polarity distribution across nodes, highlighting its role in making ECG both robust to initialization choices and scalable to large, domain-specific corpora.

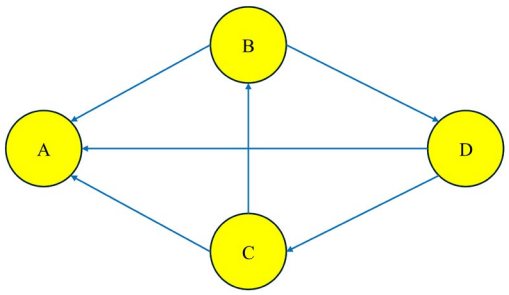

Fig. 3. PageRank algorithm.

We've observed that this weighting strategy makes a real difference during propagation. When highly connected nodes receive sentiment values, they tend to pass those values on more widely and consistently. This process helps keep the graph consistent, even when it starts with only a few seed words.

G. Challenges in Lexicon-Based Sentiment Analysis for Noun-Dominant Corpora

Lexicon-based sentiment analysis typically focuses on adjectives and adverbs due to their explicit sentiment orientation. However, real-world documents—particularly in narrative and domain-specific contexts—tend to be noun-dominant. Existing sentiment lexicons underperform in these settings because they lack sentiment values for most nouns. This presents a unique challenge when applying traditional methods to corpora such as hotel reviews or literary texts, where sentiment is often implied through noun-based co-occurrence rather than explicit sentiment terms.

To address this issue, the proposed framework focuses on building sentiment lexicons for nouns, exploring how sentiment can be propagated indirectly via co-occurrence networks without relying on predefined adjective-based polarity.

III. The Proposed Framework

The Enhanced Sentiment Analysis using Extended Corpus and Co-occurrence Graph (ECG) framework was designed to address sentiment analysis challenges in domains with limited lexical resources. This method emphasizes the role of noun-based associations within a document by leveraging their co-occurrence patterns to facilitate sentiment expansion. The process begins with a small set of sentiment-labeled seed words and incrementally propagates values to other terms through graph-based structural connections and statistical weighting, allowing for broader sentiment coverage without requiring extensive labeled data.

Unlike existing lexicon-based methods that often suffer from limited coverage, domain dependency, and static polarity assignments, ECG leverages corpus-internal co-occurrence structures to overcome these challenges. By relying on graph-based propagation instead of static lexicons, the framework provides a scalable and adaptive approach suitable for domains where high-quality lexicons are absent.

The proposed framework is composed of five sequential stages, each contributing to the systematic propagation of sentiment values throughout the document's lexical structure.

(1) Preprocessing: The process begins with text normalization and cleansing, followed by tokenization. Only noun terms are retained, as they are considered more likely to convey relevant sentiment information.

(2) Building the Co-occurrence Graph: To capture noun relationships within the text, we build a graph in which each node stands for a noun, and an edge is drawn when two nouns appear in the same sentence. The strength of each edge is measured using the Dice coefficient, along with semantic closeness.

(3) Initializing Sentiment Values: A small set of seed words, taken from sentiment resources like SentiWords, is used to assign initial polarity scores. These sentiment values are then embedded into the graph to serve as starting points for propagation.

(4) Expanding the Sentiment Corpus: Sentiment values are propagated throughout the graph using an iterative two-step approach. The process begins with immediate neighbors and proceeds to more distant nodes based on their structural connectivity.

(5) Sentiment Evaluation: After the propagation process, the finalized sentiment scores are used to classify the overall polarity of individual documents or their constituent segments.

The overall process is shown in Fig. 4. It starts from preprocessing raw text and moves through graph construction, sentiment value propagation, and finally to sentiment classification. The diagram helps visualize how each stage connects—from input documents to the final polarity decision.

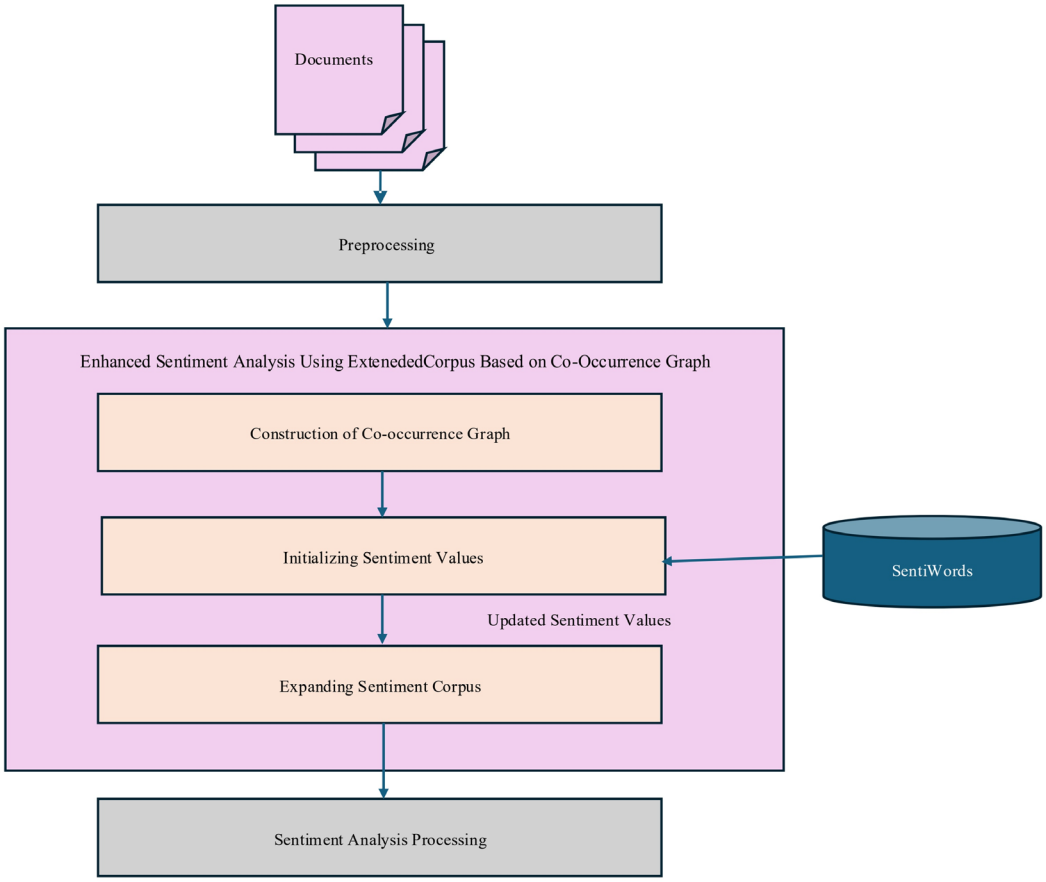

Fig. 4. Overview of the ECG framework for sentiment analysis.

As illustrated in Fig. 4, the framework introduces a feedback loop between initialization and expansion, allowing continuous refinement of sentiment values across the graph. This iterative design distinguishes ECG from traditional lexicon-based methods, which typically rely on static dictionaries. By embedding this loop, ECG ensures that polarity assignments remain consistent and extensible as the corpus grows.

A. Seed Word Initialization

The initialization phase employs nouns as the seed set, guided by two primary factors: (1) nouns frequently appear in narrative corpora and serve as stable semantic anchors (e.g., entities and locations), and (2) nouns facilitate the analysis of entity-centric sentiment evolution across chapters. Although this study focuses on nouns, future extensions aim to include adjectives and verbs to capture more expressive sentiment dimensions. Two configurations are considered: 10 positive and 10 negative seed words, and 20 positive and 20 negative seed words, all selected from SentiWords. Positive seeds are initialized with a value of +1 and negative seeds with −1, while all other nodes are initialized to 0.

Although our experiments employ the Harry Potter corpus as a narrative test case, the ECG framework is designed to generalize across multiple domains. Since sentiment propagation relies on structural relations among nouns rather than domain-specific prior lexicons, the same approach can be applied to medical texts, legal documents, or customer reviews by reinitializing the seed set. This property enables ECG to adapt effectively to low-resource or specialized fields where annotated sentiment corpora are scarce.

While our initial configuration focuses solely on nouns, this decision is grounded in both practical and theoretical considerations. Nouns tend to appear more consistently across documents and are less context-sensitive compared to adjectives and verbs, which can exhibit high variability in meaning depending on syntactic position or domain usage. Moreover, many public sentiment lexicons, such as SentiWords, contain limited coverage of sentiment-labeled adjectives and verbs in specific domains (e.g., fantasy narratives like Harry Potter). As a result, using nouns allowed us to establish a stable foundation for propagation. However, we acknowledge that nouns may not carry the full expressive range of sentiment, especially for affective or emotional states, which are often conveyed through adjectives (e.g., “happy”, “terrible”) or verbs (e.g., “love”, “hate”). In future work, we plan to integrate syntactic parsing and part-of-speech-specific weighting to include other word types, thereby enriching the sentiment dimension and improving document-level classification accuracy.

B. Data and Preprocessing

The effectiveness of the ECG framework relies on both the quality of the input corpus and the rigor of preprocessing. In this section, we first describe the dataset used in our experiments and then detail the preprocessing pipeline designed to normalize the text and extract noun-

based lexical structures that serve as the foundation for graph construction.

1) Dataset

For our experiments, we used the full set of seven novels from the *Harry Potter* series by J.K. Rowling, stored as plain text files. This corpus was selected for its rich emotional content and consistent narrative style, making it suitable for testing sentiment propagation in graph-based analysis. After preprocessing and graph construction, the corpus yielded an average of 6394 sentences per book, 3540 unique noun terms, and approximately 91,130 co-occurrence edges.

2) Preprocessing

The input text was normalized and prepared before graph construction through the following steps:

- Text Cleaning: Headers, footers, special characters, and non-informative symbols were removed. All text was converted to lowercase for consistency.
- Sentence Segmentation and Tokenization: The cleaned text was divided into sentences and further tokenized into words.
- Stopword Removal: Common function words such as the, is, and and were filtered out.
- Part-of-Speech Tagging: Each token was tagged, and only nouns and proper nouns were retained for graph construction [12].
- Noun Filtering: Sentences with fewer than two noun terms were excluded to ensure that every sentence contributed at least one valid noun pair.

This preprocessing pipeline ensured that the resulting dataset emphasized noun-based semantic relations, forming a robust structural foundation for sentiment propagation in the ECG framework.

C. Graph-Based Sentiment Corpus Expansion

The sentiment expansion process in the ECG framework consists of three stages: (i) construction of the co-occurrence graph, (ii) initialization of sentiment values, and (iii) iterative propagation of sentiment across the graph.

1) Co-occurrence graph construction

Each noun term is represented as a node, and an edge is created if two nouns co-occur within the same sentence. Formally, the co-occurrence graph G is defined in Eq. (2):

$$G = (V, E) \tag{2}$$

where G denotes the co-occurrence graph, V is the set of noun terms, and E is the set of edges representing co-occurrence relationships.

The *Dice* coefficient [13, 14] measures the strength of association between two terms A and B, as defined in Eq. (3):

$$Dice(A,B) = \frac{2\times|S_A \cap S_B|}{|S_A|\times|S_B|} \tag{3}$$

where S_A and S_B are the sets of sentences containing terms A and B.

The distance metric is defined as Eq. (4), and is used to measure the dissimilarity between two terms based on their co-occurrence strength. This metric is later used for edge pruning in the co-occurrence graph.

$$Distance(A,B) = \frac{1}{Dice(A,B)+\varepsilon} \tag{4}$$

where $\varepsilon = 0.1$ prevents division by zero.

Edges with $Dice < \theta$ are pruned to reduce noise.

The steps for generating the co-occurrence graph are summarized in Algorithm 1.

Algorithm 1: Co-Occurrence Graph Construction

Input: Corpus D, threshold θ, smoothing parameter $\varepsilon = 0.1$
Output: Graph $G = (V, E)$
1: Initialize empty graph $G = (V, E)$
2: **For** each sentence $s \in D$, segment text and apply POS tagging
3: **For** each noun pair $(t_i, t_j) \in T_s$, increment co-occurrence count
4: Compute Dice coefficient for all noun pairs
5: **If** Dice $\geq \theta$, add nodes and weighted edge to G
6: **Return** final co-occurrence graph G

2) Initialization of sentiment values

A seed set of 20 positive and 20 negative words was selected from the SentiWords lexicon.

- Positive seeds were assigned +1.
- Negative seeds were assigned −1 (words with polarity < -0.2).
- All other nodes were initialized to 0.

This provided the starting sentiment distribution for propagation.

3) Sentiment propagation

Sentiment values were expanded iteratively across the graph in three phases:

a) Phase 1—Neighbor update

Each neighbor node sentiment is updated by combining its previous value with the average of its direct neighbors, as shown in Eq. (5):

$$s(w_{ij}, t_n) = 0.8\times s(w_{ij}, t_{n-1}) + 0.2\times\left(\frac{\sum s(w_{ij}, t_{n-1})}{number\ of\ neighbor\ node(s)}\right) \tag{5}$$

In Eq. (5), $s(w_{ij}, t_n)$ denotes the updated sentiment value of node w_{ij} at iteration t_n. The term $0.8\times s(w_{ij}, t_{n-1})$ represents the contribution from the node's previous sentiment value.

The term $0.2\times\left(\frac{\sum s(w_{ij}, t_{n-1})}{number\ of\ neighbor\ node(s)}\right)$ represents the average sentiment value of its direct neighboring nodes from the previous iteration.

The coefficients 0.8 and 0.2 are weighting factors that balance the influence between the node's own history and the sentiment propagated from its neighbors.

b) Phase 2—Target node update

The target node incorporates both its previous value and the sentiment of its second-order neighbors, as shown in Eq. (6):

$$s(w_i, t_n) = 1.0\times s(w_i, t_{n-1}) + 0.25\times\left(\frac{\sum s(w_i, t_{n-1})}{number\ of\ neighbor\ node(s)}\right) \tag{6}$$

In Eq. (6), the sentiment value of the target node w_i at iteration t_n is updated by incorporating two components. The first component is the previous sentiment value $s(w_i, t_n)$, weighted by a factor of 1.0, reflecting the influence of the target node's own prior state. The second component aggregates the sentiment values of the second-order neighbors (i.e., neighbors of its direct neighbors). The average sentiment of these second-order neighbors is weighted by 0.25, indicating their reduced but meaningful influence on the target node.

Together, these two factors determine the updated sentiment score of w_i, enabling the model to integrate broader contextual sentiment information from the surrounding graph structure.

c) Phase 3—Iterative refinement

The process is repeated for up to 1500 iterations, or until convergence, as detailed in Algorithm 2. Sentiment stabilizes as values propagate across connected nodes.

Algorithm 2: Sentiment Propagation

Input: Co-occurrence graph *G* (*V*, *E*), seed sentiment values
Output: Updated sentiment values for all nodes
1. Initialize sentiment values for seed words (+1, –1); set others to 0.
2. **For** iteration = 1 to 1500 do
   For each node *wi* do
      a. Update neighbor nodes using Eq. (5)
      b. Update target node using Eq. (6)
3. Stop **if** changes fall below the convergence threshold.

To ensure stable propagation behavior, the update process was iterated up to a maximum of 1500 steps or until convergence was detected based on a fixed threshold. A PageRank-based weighting mechanism was applied to account for the influence of high-degree nodes during the propagation process.

D. Sentiment Polarity Score

After the sentiment values of all nodes in the co-occurrence graph have stabilized through iterative propagation, the final step is to determine the overall sentiment polarity of each document. This process consists of two stages: (1) calculating the mean sentiment score based on the identified nouns in the document, and (2) classifying the document sentiment using defined threshold values.

1) Stage 1: Calculating the mean sentiment score

After the sentiment values within the graph have stabilized, the next phase involves evaluating the overall sentiment of each document. This process is carried out in two steps.

First, the mean sentiment score (α) is computed by averaging the sentiment values of all noun terms present in the document, as shown in Eq. (7).

$$\alpha = \frac{1}{N} \sum_{i=1}^{N} s(w_i) \tag{7}$$

where:

- $s(w_i)$ is the sentiment value of noun w_i,
- N is the total number of nouns appearing in the document.

2) Stage 2: Determining sentiment polarity

Then, we classify the document's sentiment based on this average score, as defined in Eq. (8).

$$sentiment\ polarity = \begin{cases} positive & if\ \alpha > 0.2 \\ neutral & if -0.2 \le \alpha \le 0.2 \\ negative & if\ \alpha < -0.2 \end{cases} \tag{8}$$

- If $\alpha > 0.2$, we label it as positive.
- If $\alpha < -0.2$, we label it as negative.
- If $-0.2 \le \alpha \le 0.2$, we label it as neutral.

Then, to classify the document's sentiment based on this average score, we apply a set of threshold values. After testing different ranges, we found that ±0.2 worked best for capturing both strong and more subtle emotional changes, especially in literary texts like novels. This evaluation method relies on two steps: averaging the sentiment values of nouns and using clear cut-offs to decide the overall tone. With this approach, we can assess a document's sentiment even when only a few seed words are available for training.

IV. Experiment and Results

This section describes how the ECG framework was evaluated in practice. We detail the dataset, experimental design, evaluation metrics, and results. The objective was to examine whether sentiment values derived from a limited number of seed words could effectively propagate across a narrative corpus, and to compare the method against conventional lexicon-based sentiment analysis.

A. Experimental Setup

The analysis was conducted on the full set of seven novels from the *Harry Potter* series authored by J.K. Rowling. This corpus was selected for three reasons:

(1) It contains a rich and consistent emotional tone across multiple chapters.
(2) It includes numerous entities (characters, places, objects) that provide fertile ground for sentiment propagation.
(3) Its narrative vocabulary is domain-specific and often absent from standard sentiment lexicons such as SentiWordNet and VADER.

Before graph construction, the text underwent preprocessing as outlined in Section III, including sentence segmentation, tokenization, stopword removal, and part-of-speech tagging. Only nouns and proper nouns were retained, and sentences with fewer than two nouns were discarded to guarantee valid co-occurrence pairs.

A co-occurrence graph was then built for each chapter: nouns became nodes, and edges represented co-occurrence within the same sentence. Edge weights were computed using the Dice coefficient. To mitigate noise, a threshold was applied so that only meaningful connections remained.

For Seed Word Configurations, we tested Two experimental configurations:

- Condition A: 10 positive and 10 negative seed words
- Condition B: 20 positive and 20 negative seed words

Seed terms were drawn from the SentiWords lexicon, prioritizing items with extreme polarity scores. Positive seeds (polarity > +0.4) were assigned +1, negative seeds

(polarity < –0.4) were assigned –1, and all other nodes initialized to 0. This setup reflects a low-resource scenario with minimal supervision. Sentiment propagation was iterated for 1500 rounds until convergence.

B. Experimental Design

The sentiment propagation procedure followed the three-phase algorithm detailed in Section III:

(1) Neighbor Update: Estimate the sentiment of each neighbor node using weighted values from its adjacent nodes (Eq. (5)).
(2) Target Node Update: Recalculate the sentiment of the target node by aggregating both direct and second-order neighbors (Eq. (6)).
(3) Iterative Refinement: Repeat the process over 1500 iterations or until values stabilize.

Although the majority of nodes begin as neutral, recursive updates combined with structural weighting (PageRank and triadic closure) enable sentiment information to flow gradually across the network.

C. Evaluation Metrics

The Coefficient of Variation (COV), defined in Eq. (9), was employed as the primary evaluation metric. Unlike accuracy or F1—which require labeled datasets—COV captures how sentiment scores distribute across the lexical space, offering insight into granularity and contrast.

$$COV = \left(\frac{\sigma}{\mu}\right) \times 100 \tag{9}$$

where σ is the standard deviation and μ is the mean sentiment score.

A higher COV indicates a more distinct spread of sentiment across terms, while a low COV suggests homogenization. This makes COV particularly suitable for lexicon expansion tasks where external gold labels are scarce.

To provide additional context, we also compared ECG against baseline lexicons (SentiWordNet, VADER, NRC). Comparisons focused on coverage rate (percentage of narrative nouns assigned sentiment) and consistency with narrative tone.

D. Results

The co-occurrence graph constructed from the *Harry Potter* corpus contained approximately 3540 noun nodes and about 91,130 edges per book. Sentiment propagation, initialized with either 10, 20 or 30 seed terms depending on the configuration, consistently converged within 1500 iterations. The following subsections summarize the key experimental findings.

(1) Growth of Sentiment Nodes. Fig. 5 illustrates the expansion of positive and negative nodes during propagation. Positive sentiment spread rapidly during the first 300–400 iterations before reaching a plateau around 1000 iterations, while negative sentiment grew at a slower yet steadier rate, stabilizing at a similar point. This pattern demonstrates that the propagation algorithm achieved stable convergence, with neutral nodes gradually transforming into either positive or negative sentiment categories. Such stability is particularly important in low-resource scenarios, where only a minimal set of seed words is available. Even with limited supervision, the ECG framework effectively distributed sentiment values across thousands of narrative terms, addressing the challenge of sentiment analysis in domains with scarce annotated resources, as discussed in Section I.

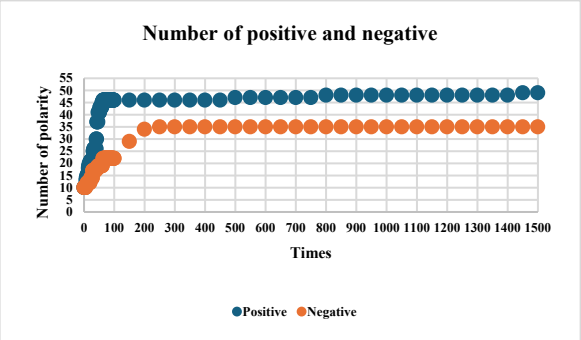

Fig. 5. Growth in the number of sentiment-assigned words over 1500 iterations.

(2) Coefficient of Variation (COV). Table I summarizes the Coefficient of Variation (COV) values obtained under different seed configurations. Expanding the seed set from 10 to 20 words consistently improved the COV, with the highest value (55.75%) observed under the Positive & Negative condition using 20 seeds. Fig. 6 provides a visual comparison, illustrating that richer seed supervision produces sharper sentiment differentiation across the graph. These results demonstrate that a larger and more balanced seed set enhances the model's ability to capture subtle emotional variations. The increase in COV with larger seed sets not only indicates improved differentiation of polarity values but also reflects enhanced generalizability. A higher COV represents a lexicon space in which sentiment values are more finely distributed across diverse terms, allowing the model to adapt more effectively to unseen or domain-specific vocabulary. Overall, these findings highlight the ECG framework's potential for robust cross-domain sentiment analysis.

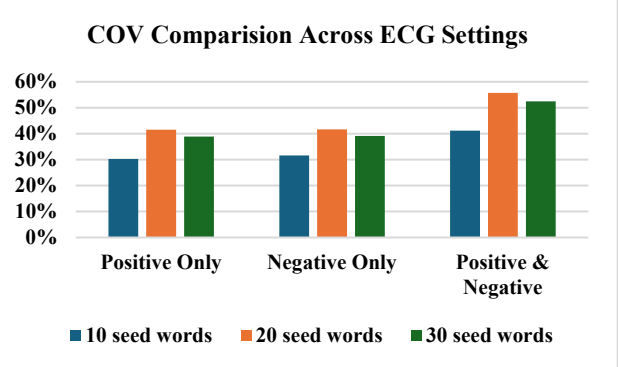

Fig. 6. COV comparison across polarity settings under 10, 20 and 30 seed word configurations.

(3) Baseline Comparison. To further contextualize ECG's performance, we compared its coverage against common lexicon-based approaches. SentiWordNet covered only 18% of narrative nouns, VADER reached

12%, and NRC Emotion Lexicon improved to 21%. However, these lexicons consistently failed to assign sentiment to domain-specific terms critical to narrative interpretation (e.g., "Slytherin", "broomstick", "Patronus"). Their static assignment of polarity values limited adaptability, resulting in mismatches with narrative tone across chapters. In contrast, ECG achieved 100% coverage of retained nouns, leveraging graph-based propagation to integrate context-specific sentiment dynamically. This result highlights ECG's capacity for domain adaptation, a key limitation of conventional lexicons.SentiWordNet covered only 18% of narrative nouns.

- VADER reached 12% coverage, but failed to capture sentiment of domain-specific terms *(*e.g., Slytherin, broomstick*)*.
- NRC Emotion Lexicon improved to 21% coverage but exhibited inconsistency across chapters.
- In contrast, ECG achieved 100% coverage of retained nouns, with sentiment assignments aligning coherently with narrative tone.

This comparison demonstrates ECG's ability to generalize beyond the limitations of fixed lexicons, particularly in domain-rich narrative corpora. Moreover, the adaptive nature of ECG enables continuous sentiment refinement as new contextual data become available, making it well-suited for dynamic or evolving text sources. In practical applications, this adaptability allows the framework to maintain consistency across diverse genres and languages without requiring manual lexicon reconstruction. Detailed sentiment values and coverage comparison are reported in Appendix (Table A1 and Table A2).

These results are consistent with the trends observed during sentiment propagation. Increasing the number of seed words from 10 to 20 consistently improved the COV across all test conditions. The highest COV (55.75%) was obtained under the Positive & Negative configuration using 20 seed words. This indicates that sentiment values were more widely and distinctly distributed throughout the graph, enhancing the model's ability to differentiate between emotional tones.

To further examine whether adding additional seed words would enhance performance, an experiment was conducted with 30 seed words. The results showed that the COV slightly increased compared with the 20-seed configuration across all conditions. As presented in Table I, this suggests that while expanding the seed set improves performance up to a certain threshold, excessive initialization (e.g., 30 words) may introduce semantic overlap or noise that reduces propagation stability. Therefore, the 20-seed configuration provides the optimal balance between coverage and clarity.

Fig. 6 illustrates the comparison of Coefficient of Variation (COV) across three different seed word configurations (10, 20, and 30) under three polarity conditions: Positive Only, Negative Only, and Positive & Negative.

The results indicate that:

- The 10-seed configuration consistently yields the lowest COV values, suggesting the most stable sentiment propagation.
- Increasing seed words to 20 improves differentiation but introduces greater variation, especially when both positive and negative words are used.
- Interestingly, 30 seed words reduce variation compared to 20, yet not as low as the 10-seed case—indicating that adding more seeds beyond a point may introduce semantic noise or redundancy. Overall, these findings emphasize the trade-off between seed set size and propagation stability, indicating that an intermediate configuration (around 20 seeds) provides the optimal balance for the ECG framework.

E. Discussion

The experimental results provide three key insights into the effectiveness and adaptability of the ECG framework:

(1) Low-Resource Viability. The ECG framework demonstrated that even with only ten seed words, it can achieve meaningful sentiment propagation across a large lexical space. This result underscores its suitability for low-resource domains where annotated sentiment lexicons are limited or unavailable.

(2) Improved Granularity. Expanding the seed set to twenty terms significantly enhanced sentiment differentiation, as evidenced by higher COV values. This finding indicates that richer supervision enables the framework to generate more fine-grained sentiment distributions, effectively capturing subtle emotional variations across chapters.

(3) Compared with SentiWordNet, VADER, and the NRC Emotion Lexicon, the ECG framework demonstrated broader coverage and more context-aware sentiment assignments. Domain- and narrative-specific terms that conventional lexicons failed to capture were effectively incorporated into ECG's sentiment representation, producing outputs closely aligned with each narrative's contextual flow.

(4) One of the key design decisions in this study was to initialize sentiment propagation using nouns as seed words. This choice was motivated by two rationales. First, nouns constitute the most abundant and domain-representative part of speech in the selected corpus (i.e., literary narratives), encompassing characters, entities, and key plot elements. Second, unlike adjectives or adverbs whose sentiment values are readily available in existing lexicons, nouns generally lack sentiment annotations—presenting a unique challenge for propagation-based sentiment expansion. By focusing on nouns, the proposed method addresses a critical gap in sentiment analysis research: extending sentiment knowledge to parts of speech that are underrepresented in conventional lexicons such as SentiWordNet and VADER.

Furthermore, the exclusive focus on nouns for both seed initialization and lexicon expansion represents a deliberate methodological choice to examine sentiment propagation within the most challenging part-of-speech category. Unlike adjectives and adverbs—whose sentiment polarity is explicitly expressed (e.g., happy, terrible)—nouns

typically convey latent sentiment that emerges through contextual associations (e.g., funeral, celebration). Expanding sentiment over nouns not only bridges a major gap in existing lexicons but also facilitates context-aware sentiment analysis, which is particularly critical in specialized domains such as healthcare and legal texts, where nouns often carry substantial emotional and decision-making implications. This design therefore provides a rigorous test of the propagation algorithm's capability under minimal lexical supervision.

Additionally, in domains such as medical or technical documents, key decision-related terms are predominantly noun-based (e.g., tumor, procedure, crisis, recovery) rather than purely adjectival. Thus, focusing on nouns represents both a methodological challenge and an essential step toward broader domain generalization. Future work may extend this approach by incorporating other parts of speech (e.g., adjectives and verbs) to further enrich sentiment propagation beyond nouns, as discussed earlier.

Scalability and Practicality. Executing 1500 iterations on the full Harry Potter corpus (approximately 1.1 million tokens) required around 25 min on a standard workstation equipped with an Intel i7 processor and 16 GB of Random Access Memory (RAM), while maintaining memory usage below 4 GB. The iterative framework is inherently scalable to larger corpora when combined with graph pruning strategies, thereby ensuring computational efficiency and feasibility for real-world applications.

Beyond scalability, the ECG framework also demonstrates strong generalizability. Although this study focused on narrative text, the framework is not confined to literary domains. Its reliance on co-occurrence structures and propagation dynamics—specifically PageRank weighting and Triadic Closure—enables adaptation across diverse genres such as medical case reports, legal documents, and social media discourse, where conventional lexicons often struggle due to limited vocabulary coverage. By requiring only a small set of initial seed words, ECG offers a practical and efficient pathway for sentiment analysis in resource-scarce domains, thereby underscoring its extensibility beyond the current corpus.

Although the experiments in this study focused on narrative texts, the ECG framework demonstrates adaptability to other genres and languages. Future extensions may integrate adjectives and verbs (see Section III) to enhance sentiment dimensionality. Moreover, the framework holds promise for applications across multilingual and domain-specific corpora—such as medical, legal, and technical texts—where conventional lexicons often fall short.

V. Conclusion

This study introduced the Enhanced Sentiment Analysis Using Extended Corpus and Co-occurrence Graph (ECG) framework for expanding sentiment polarity from a limited set of seed nouns. By leveraging graph-based representations and iterative propagation, ECG successfully assigned sentiment values to unlabeled terms within narrative texts. The experimental evaluation confirmed that ECG can effectively infer sentiment polarity from minimal supervision while maintaining full lexical coverage in the target corpus. The framework contributes a scalable and interpretable approach to sentiment expansion, offering improved applicability in domains where sentiment annotations are scarce.

Appendix: Supplementary Data for Sentiment Analysis

TABLE A1. Sentiment Values of Selected Words After 1500 Iterations of Propagation

Word	Sentiment at t = 0	Sentiment at t = 1500
Positive	1.00	49.00
Negative	−1.00	35.00
team	0.00	1.00
broomstick	0.00	1.00
slytherin	0.00	1.00
mirror	0.00	−0.95
face	0.00	−0.98
neck	0.00	−0.99
snake	0.00	−0.99
voldemort	0.00	0.00

TABLE A2. Sentiment Coverage and Consistency: ECG vs Baseline Lexicons

Method	Coverage (%)	Domain-Specific Handling	Consistency Across Chapters
SentiWordNet	18	Poor (missed narrative terms)	Moderate
VADER	12	Very Poor (fails on fantasy terms)	Low
NRC Lexicon	21	Limited, partial coverage	Inconsistent
ECG (Ours)	100	Strong (handles "Slytherin", "broomstick")	High

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

NR conducted the research, implemented the co-occurrence graph and sentiment propagation framework, and wrote the manuscript. MM contributed to the result analysis and provided feedback on the interpretation of findings. HU proposed the research direction and provided supervision throughout the project. All authors reviewed and approved the final version of the manuscript.

Acknowledgment

The author wishes to thank Asst. Prof. Dr. Maleerat Maliyaem for her valuable suggestions on the interpretation of results, and Prof. Dr.-Ing. habil. Dr. h.c. Herwig Unger for his insightful guidance in shaping the research direction. Their support and supervision throughout the research process were greatly appreciated. The author also appreciates the constructive feedback received during the manuscript revision, which helped improve the clarity and coherence of this study.

REFERENCES

[1] H. Liu, I. Chatterjee, M. Zhou *et al.*, "Aspect-based sentiment analysis: A survey of deep learning methods," *IEEE Transactions on Computational Social Systems,* vol. 7, no. 6, pp. 1358–1375, 2020.

[2] B. Kumar, Sheetal, V. S. Badiger *et al.*, "Sentiment analysis for products review based on NLP using lexicon-based approach and roberta," in *Proc. 2024 International Conf. on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, 2024, pp. 1–6.

[3] P. K. Punde and R. S. Wagh, "A survey paper on different approaches for sentiment analysis," in *Proc. 2018 Second International Conf. on Electronics, Communication and Aerospace Technology (ICECA*), 2018, pp. 203–207.

[4] T. R. N and R. Gupta, "A survey on machine learning approaches and its techniques," in *Proc. 2020 IEEE International Students' Conf. on Electrical, Electronics and Computer Science (SCEECS)*, 2020, pp. 1–6.

[5] M. Yang, X. Chen, M. Zhao *et al.*, "Dynamic social network embedding based on triadic closure pattern analysis," in *Proc. 2021 20th International Conf. on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, 2021, pp. 302–308.

[6] F. A. Nugraha and Z. K. A. Baizal, "Ontology-based recommender system for bandung region eateries with sentiment analysis enhancement," in *Proc. 2024 IEEE International Conf. on Communication, Networks and Satellite (COMNETSAT)*, 2024, pp. 109–114.

[7] Z. Zhu, Z. Qiu, and Y. Zhao, "Importance identification of distribution network nodes based on improved pagerank algorithm," in *Proc. 2022 First International Conf. on Cyber-Energy Systems and Intelligent Energy (ICCSIE)*, 2023, pp. 1–6.

[8] S. Ahmadian and S. Haddadan, "A theoretical analysis of graph evolution caused by triadic closure and algorithmic implications," in *Proc. 2020 IEEE International Conf. on Big Data (Big Data)*, 2020, pp. 5–14.

[9] S. Park, W. Lee, B. Choe *et al.*, "A Survey on personalized pagerank computation algorithms," *IEEE Access,* vol. 7, pp. 163049–163062, 2019.

[10] S. Simcharoen and H. Unger, "Graph-based word clustering considering the distance and the connectivity of a co-occurrence," in *Proc. 2022 Research, Invention, and Innovation Congress: Innovative Electricals and Electronics (RI2C)*, 2022, pp. 238–242.

[11] M. Zhu, M. Li, Z. Geng *et al.*, "Dice coefficient matching-based sparsity adaptive matching pursuit algorithm for the digital predistortion model pruning," in *Proc. 2018 IEEE 18th International Conf. on Communication Technology (ICCT)*, 2018, pp. 1032–1035.

[12] H. Zhang and M. O. Shafiq, "Survey of transformers and towards ensemble learning using transformers for natural language processing," *Journal of Big Data,* vol. 11, 25, 2024.

[13] A. W. Setiawan, "Image segmentation metrics in skin lesion: Accuracy, sensitivity, specificity, dice coefficient, jaccard index, and matthews correlation coefficient," in *Proc. 2020 International Conf. on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, 2020, pp. 97–102.

[14] A. Xie, E. Elfatimi, S. Ghosal *et al.*, "Deep learning with uncertainty quantification for predicting the segmentation dice coefficient of prostate cancer biopsy images," in *Proc. 2024 International Conf. on Machine Learning and Applications (ICMLA)*, 2024, pp. 1158–1163.

[15] S. H. Chiu and B. Chen, "Innovative bert-based reranking language models for speech recognition," in *Proc. 2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 266–271.

[16] S. Supal, S. M. Anzar, C. Jacob *et al.*, "Deep learning transformers for sentiment classification: A performance evaluation," in *Proc. 6th International Conf. on Control, Communication and Computing (ICCC)*, 2025, pp. 1–6.

[17] P. Porkaew, P. Boonkwan, and T. Supnithi, "HoogBERTa: Multi-task sequence labeling using thai pretrained language representation," in *Proc. 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2021, pp. 1–6.

[18] C. Huang, Y. Luo, and Q. Li, "Text classification method based on co-occurrence events," in *Proc. 2019 15th International Conf. on Computational Intelligence and Security (CIS)*, 2019, pp. 277–281.

[19] O. Leila and B. Djamal, "A vector space based approach for short answer grading system," in *Proc. 2018 International Arab Conf. on Information Technology (ACIT)*, 2018, pp. 1–9.

[20] R. Zhang, D. S. Lee, and C. S. Chang, "A generalized configuration model with triadic closure," *IEEE Transactions on Network Science and Engineering,* vol. 10, no. 2, pp. 754–765, 2023.

[21] R. Sardhara and K. I. Lakhataria, "An improved pagerank algorithm based on reachability to reduce mutual reinforcement effect," in *Proc. 2019 10th International Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–6.