International Journal of Analysis and Applications

Optimizing Deep Learning-Based Hate Speech Classification on Social Media Through Frame Semantic Theory and Hyperparameter Tuning

Korakoch Mounggam, Kittipol Wisaeng*

Mahasarakham Business School, Mahasarakham University, Mahasarakham 44150, Thailand
**Corresponding author:* Kittipol.w@acc.msu.ac.th

ABSTRACT. This study uses deep learning, frame semantic theory, and precise hyperparameter tuning to classify social media hate speech. Three carefully selected hate speech datasets, arranged according to a frame semantic framework, are used to categorize opposing content, helping to better grasp context than simply looking for words. The Synthetic Minority Over-sampling Technique (SMOTE) enhanced minority-class detection to mitigate class imbalance. To find optimal configurations, deep learning architectures, activation functions, and data split algorithms were tested. The best approach used 70% of the data for training and 30% for testing, a model with hidden layers, a Rectifier activation function, and 10 epochs, achieving 96.30% accuracy, 96.95% recall, 99.15% precision, and 0.98 F1-score. These findings show that frame semantic theory, deep learning, and rigorous hyperparameter optimization can significantly improve social media hate speech detection systems. The method lays the foundation for context-aware and socially responsible content control.

1. Introduction

The rapid proliferation of social media platforms has fundamentally transformed global communication, breaking down geographical boundaries, enabling instantaneous information exchange, and fostering diverse online communities that span cultural, linguistic, and ideological divides. Social networks such as Facebook, Twitter (now X), Instagram, and TikTok have empowered individuals to share opinions, mobilize collective action, and engage in meaningful dialogue with unprecedented reach and immediacy. However, this openness and accessibility have also created fertile ground for the spread of harmful and divisive content, particularly in the form of hate speech. Hate speech poses serious risks to personal well-being, societal cohesion,

Received Aug. 16, 2025

2020 Mathematics Subject Classification. 68T07.

Key words and phrases. deep learning; frame semantics; hate speech classification; hyperparameter tuning.

https://doi.org/10.28924/2291-8639-24-2026-86

ISSN: 2291-8639

and democratic discourse. In recent years, the escalation of such content has underscored the urgent need for effective detection and mitigation strategies to protect vulnerable communities and preserve the integrity of online interactions [1]. Hate speech detection has thus emerged as a crucial area of research within computational linguistics, natural language processing (NLP), and artificial intelligence (AI). Conventional detection approaches have typically relied on keyword-based filtering or traditional machine learning models such as Support Vector Machines (SVM) and Naïve Bayes. While these methods are relatively straightforward to implement, they often fail to account for the nuanced and context-dependent nature of language. Words or phrases that are benign in one context may be harmful in another, and the same hateful intent can be expressed in varied linguistic forms. As a result, conventional methods tend to suffer from high false-positive and false-negative rates, reducing their trustworthiness and limiting their utility for real-world deployment [2]. In contrast, deep learning offers a powerful alternative by modeling complex, nonlinear relationships and capturing rich semantic representations from large volumes of text. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based models (e.g., BERT, RoBERTa) have demonstrated superior performance in various NLP tasks, including sentiment analysis, offensive language detection, and contextual text classification. These models excel at capturing word dependencies, context, and subtle variations in tone, making them well-suited to address the inherent challenges in detecting hate speech. However, even deep learning approaches can struggle with semantic ambiguity, implicit hate speech, and the cultural or domain-specific interpretations of language.

Recent developments in linguistic theory, particularly frame semantic theory, offer promising avenues for addressing these limitations. Frame semantics posits that the meaning of a word is understood within a broader conceptual structure or "frame" that represents the background knowledge and relationships associated with that word [3], [4]. For example, the meaning of the word "attack" is not fully interpretable without understanding the entities involved, their roles, and the situational context. Applying frame semantics to hate speech detection enables the classification process to move beyond surface-level lexical cues and focus instead on the underlying concepts, intentions, and relational structures embedded in the text. This approach enables more accurate differentiation between innocuous expressions and genuinely harmful discourse, thereby reducing the misclassification rates that plague simpler methods. The integration of frame semantic theory with deep learning-based architectures holds the potential to yield models that are both context-aware and semantically informed. By embedding linguistic frames into neural architectures, models can be trained to detect not only overt hate speech but also more implicit, coded, or metaphorical forms of harmful language. Furthermore, to address the frequent class imbalance in hate speech datasets—where non-hate

speech examples typically outnumber hate speech examples by a large margin—this study incorporates the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic training samples for minority classes, enabling the model to learn more balanced decision boundaries and improving recall for underrepresented categories without sacrificing precision.

The present research aims to systematically optimize deep learning-based hate speech classification on social media by combining the theoretical rigor of frame semantics with the empirical advantages of hyperparameter tuning and balanced data sampling. Multiple model architectures and parameter configurations will be explored to identify the optimal design for both accuracy and reliability. The experimental framework leverages labeled datasets that are enriched with semantic annotations, ensuring that the models are trained to recognize not only the presence of specific words but also the contextual roles those words play in expressing harmful intent. By combining advanced computational methods with linguistic insights, the proposed approach aims to enhance the precision, robustness, and ethical foundation of automated hate speech detection systems. Beyond immediate technical improvements, this work has broader implications for the development of responsible AI in online content moderation. More nuanced and context-aware detection models can help social media platforms foster safer, more inclusive digital spaces, support constructive public discourse, and mitigate the harmful societal impacts of online hate. The findings from this study could thus pave the way for the next generation of intelligent moderation tools, systems that not only identify hateful content with greater accuracy but also interpret its motivations and conceptual underpinnings, enabling more informed, transparent, and equitable content governance practices.

2. Literature review

The Challenge of Hate Speech Detection on Social Media

The rapid expansion of user-generated content (UGC) on social media platforms has fundamentally reshaped digital communication while simultaneously introducing unprecedented challenges for content moderation. Platforms such as Facebook, X (formerly Twitter), YouTube, and TikTok host billions of daily interactions, many of which contain diverse linguistic expressions, cultural references, and emergent trends. This immense volume and diversity of content make identifying and moderating harmful language exceptionally complex [5]. Hate speech poses a unique detection challenge due to its context-dependent nature, the diversity of its manifestations, and its continual evolution in response to socio-political events, technological changes, and platform moderation policies [6], [7]. The dynamic interplay between evolving language use and user attempts to evade detection further complicates the creation of reliable automated moderation systems. Traditional keyword-based moderation methods, while

straightforward and computationally efficient, remain inadequate in addressing these complexities. Such approaches often operate by matching text against a precompiled list of offensive terms. However, they are inherently limited in their ability to detect implicit hate speech, sarcasm, metaphors, and other forms of indirect, harmful expressions [8], [9]. These limitations often result in high false-negative rates, as implicit messages slip past filters, and high false-positive rates, as benign uses of flagged terms are mistakenly categorized as harmful. The resulting inaccuracies diminish the trustworthiness of automated moderation systems and raise concerns about fairness, bias, and the over-suppression of legitimate discourse. The temporal dimension of online discourse further amplifies the detection challenge. Language is not static; its meanings, connotations, and societal implications shift over time in response to cultural movements, political developments, and technological innovations [6]. For example, once neutral terms can acquire derogatory meanings, while formerly offensive language may be reclaimed by particular communities. Consequently, hate speech detection systems must adapt to evolving linguistic landscapes in near real-time to maintain effectiveness and relevance. Recent advancements in large language models (LLMs) such as GPT, LLaMA, and PaLM have opened promising opportunities to enhance hate speech detection through improved contextual sensitivity and semantic understanding [10]. Prompt engineering has emerged as a particularly innovative technique for leveraging LLMs' reasoning capabilities. These approaches enable systems to interpret nuanced linguistic patterns, detect coded or metaphorical expressions, and adapt to new forms of hateful discourse more effectively than conventional pipelines [8]. Nevertheless, integrating LLM-based systems into moderation workflows presents its challenges, including computational costs, explainability concerns, and the risk of perpetuating biases inherent in the training data. Moreover, parallels can be drawn from security-focused domains such as cyberattack detection, where adaptive machine learning approaches are employed to identify and counteract rapidly evolving threats [9]. In these domains, resilience and adaptability are critical, as adversaries constantly modify their tactics to evade detection. Similarly, in the context of social media moderation, hate speech detection systems must be designed to update dynamically, incorporate feedback, and learn from newly identified patterns to remain robust against emerging linguistic strategies. These insights suggest that a practical approach to hate speech detection requires a hybrid of advanced computational techniques, linguistic theory, and adaptive learning mechanisms that balance precision and agility in an ever-changing online environment.

Advances in Deep Learning for Text Classification

Deep learning approaches have emerged as a cornerstone of modern natural language processing (NLP), driving substantial advances across a wide range of applications. Architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and

transformer-based models have demonstrated superior performance in tasks including text classification, sentiment analysis, document summarization, and hate speech detection [11], [12]. Unlike conventional machine learning methods, which typically rely on handcrafted features and shallow learning paradigms, deep learning models can autonomously learn and represent intricate hierarchical patterns, contextual dependencies, and latent semantic relationships directly from raw text data. This ability to extract multi-level features has led to marked improvements in classification accuracy, robustness, and adaptability in real-world scenarios [13], [14]. Recent surveys and comparative evaluations have consistently reported that deep learning architectures outperform traditional algorithms across diverse NLP tasks and domains [15]–[18]. For instance, in document analysis, CNN-based models excel at identifying discriminative n-gram patterns, while recurrent architectures such as long short-term memory (LSTM) and gated recurrent units (GRUs) effectively model sequential dependencies in language. More recently, hierarchical deep learning models (HDLTex) have enabled multi-tiered text categorization, allowing large-scale corpora to be classified at multiple levels of granularity while maintaining high accuracy. In parallel, transformer-based models such as BERT, RoBERTa, and XLNet have redefined contextual understanding in NLP by employing self-attention mechanisms to capture bidirectional dependencies and nuanced semantic relationships across entire sequences [11], [12], [19]. These innovations have been particularly transformative in tasks where subtle context and polysemy play a critical role, such as hate speech identification and offensive language detection. In addition to architectural advances, optimization techniques have played a crucial role in improving deep learning performance. Hyperparameter tuning strategies have been instrumental in identifying optimal configurations of learning rates, batch sizes, activation functions, and network depths. More sophisticated metaheuristic optimization methods, such as the Symbiotic Organisms Search (SOS) algorithm, have been successfully applied to enhance the learning efficiency of LSTM networks, enabling them to converge more rapidly while avoiding local minima [20]. Such optimizations not only improve performance metrics, such as accuracy and F1 score, but also enhance model generalization and stability across datasets. Notably, the versatility of deep learning has expanded beyond traditional NLP domains into security-critical applications, where adaptability and resilience are crucial. In fields such as cyberattack detection, financial fraud detection, and network intrusion prevention, deep learning models have demonstrated the ability to detect novel, previously unseen patterns in real-time [21]. These successes highlight the potential of deep learning as a unifying framework for handling complex, dynamic classification problems in both linguistic and non-linguistic domains. When applied to hate speech detection, these same principles can be leveraged to create systems that not only recognize explicit harmful content but also adapt to evolving linguistic patterns, emerging threats, and adversarial attempts to bypass moderation mechanisms.

The Role of Frame Semantics in Understanding Language

Frame semantic theory, first introduced by Fillmore [3], offers a foundational approach for understanding the deeper conceptual structures, referred to as "frames", that underpin the meanings of words and expressions. A frame can be thought of as a structured mental representation that encapsulates relationships among concepts, events, participants, and contextual elements associated with a given term or expression. This perspective moves beyond surface-level lexical analysis, emphasizing that meaning is not confined to individual words but emerges from their roles within broader conceptual and situational contexts. By mapping language onto these conceptual frames, frame semantics provides a systematic framework for interpreting not only what is said, but also the underlying intent, relationships, and implications embedded in the discourse. In the domain of hate speech identification, frame semantics offers a particularly valuable lens for analysis. Hate speech often manifests in subtle, indirect, or context-dependent ways, making it difficult to detect with simple keyword matching or shallow statistical methods. For instance, a phrase that appears neutral in isolation may carry derogatory or exclusionary implications when understood within the relevant cultural, political, or historical frame. By grounding analysis in conceptual frames, it becomes possible to classify expressions according to their deeper meanings, social functions, and relational roles, rather than relying solely on overt linguistic markers. This capability is significant in detecting implicit hate speech, coded language, and metaphorical constructions designed to evade keyword-based detection systems. Recent research supports the value of integrating frame semantics into natural language processing (NLP) models for improved interpretability and classification performance [22], [23]. Baker [22], for example, introduced FrameNet as a comprehensive lexical and semantic resource that systematically organizes English vocabulary into frame structures, enabling NLP systems to identify and exploit frame-level knowledge for advanced language understanding. FrameNet's annotated corpus and frame-based lexicon have been widely adopted in NLP applications, providing structured mappings between words, their meanings, and the situations they evoke. Similarly, Zymovets [23] demonstrated how frame semantics can serve as an explanatory model for the formation of commercial names, underscoring its ability to elucidate conceptual relationships and intended meanings across diverse linguistic contexts. These studies highlight the flexibility and depth of frame semantics as both a linguistic theory and a computational resource. Applying frame semantics to hate speech detection enables models to better account for the implicit, contextually anchored nature of harmful language. By incorporating frame-based features into deep learning architectures, classification systems can more accurately discern the social and cognitive functions of potentially harmful expressions. This integration not only improves detection accuracy but also enhances the interpretability of model decisions, as frame-based reasoning provides a transparent rationale for why a given expression is classified as

hateful. For instance, instead of flagging a term solely because it appears in a blacklist, a frame-informed system might identify that the term is part of a "Conflict" or "Disparagement" frame targeting a protected group, thus offering a context-sensitive explanation. Ultimately, the fusion of frame semantics with deep learning has the potential to shift hate speech detection from a surface-level, reactive process to a more nuanced, contextually aware, and explanatory approach. Such integration addresses critical challenges in automated moderation while aligning with ethical imperatives for transparency and fairness in AI-driven decision-making.

Addressing Class Imbalance in Hate Speech Datasets

Hate speech datasets often exhibit substantial class imbalance, with instances of non-hate content far outnumbering those containing explicit or implicit hate speech [24]. This imbalance poses significant challenges for accurate classification and equitable performance evaluation. Classifiers trained on such skewed datasets tend to favor the majority class, leading to disproportionately high accuracy but poor recall for the minority (hate speech) class. This imbalance not only hampers the system's ability to detect harmful content but also increases the likelihood of false negatives, potentially undermining trust in automated moderation systems and enabling the persistence of abusive discourse. Addressing class imbalance is thus a critical prerequisite for developing reliable and fair hate speech detection models. Oversampling strategies, particularly the Synthetic Minority Over-sampling Technique (SMOTE), have been widely recognized for their effectiveness in mitigating class imbalance in machine learning tasks [24]. SMOTE synthesizes new, artificial examples of minority-class samples by interpolating between existing instances, thereby enriching the diversity of the minority class without simply duplicating data. This approach enables classifiers to learn more robust and generalizable decision boundaries, thereby improving sensitivity to minority-class patterns and reducing bias toward majority-class predictions. In the context of hate speech detection, SMOTE and related methods can substantially improve the identification of subtle or infrequent hate expressions, especially those underrepresented in training datasets because explicit incidents are relatively rare compared to the abundance of neutral or non-hateful discourse. Beyond online content moderation, numerous studies have demonstrated the value of advanced oversampling and data augmentation methods for addressing challenges posed by limited or imbalanced datasets. Dou et al. [25], for example, examined a range of machine learning approaches to address small-data challenges in molecular science, emphasizing the role of synthetic data generation in improving model predictive performance. Similarly, Li et al. [26] reported that balancing rare samples via targeted oversampling improved property-prediction accuracy in solid-state electrolyte research, underscoring the cross-domain utility of these techniques. Kulichenko et al. [27] explored data production methodologies for training machine-learning-based interatomic potentials, highlighting how rigorous data preparation directly impacts the reliability of scientific modeling.

In another domain, Zhumabekova et al. [28] demonstrated that addressing data imbalance was crucial to the effectiveness of machine and deep learning models for identifying malware risks in healthcare systems. Collectively, these examples reaffirm the centrality of balanced data representation in developing robust, fair, and trustworthy machine learning applications. While both deep learning approaches and semantic analysis frameworks, such as frame semantics, have individually shown substantial promise for improving hate speech detection, research integrating these two paradigms remains limited. Deep learning excels at modeling complex hierarchical and contextual relationships in text, while frame semantics provides a rich theoretical foundation for understanding the conceptual and situational structures underlying language use. The synergy between these methods offers the potential to capture not only the lexical and syntactic features of hate speech but also the deeper conceptual roles and intentions embedded within it, particularly valuable for detecting implicit, coded, or context-dependent expressions. This study addresses this gap by proposing a novel integration of frame semantic theory into deep learning architectures, augmented by systematic hyperparameter optimization and advanced oversampling strategies such as SMOTE. The combination of semantic knowledge representation and adaptive computational modeling aims to produce classification systems that are both context-aware and resilient to class imbalance. By embedding frame-based conceptual structures into neural networks and balancing the training data to ensure fair representation of all classes, the proposed approach aims to improve precision, recall, and interpretability in hate speech classification. Ultimately, this research aims to contribute not only to the technical advancement of automated content moderation systems but also to the development of ethically grounded, transparent, and socially responsible AI solutions that foster safer, more inclusive online environments.

The Purpose of the Study

This study seeks to advance the field of automated hate speech detection by integrating frame semantic theory into deep learning-based classification frameworks, thereby enabling models to perform more context-aware, semantically informed, and conceptually grounded categorization of hateful language. By leveraging the theoretical depth of frame semantics, the proposed approach aims to move beyond surface-level lexical matching and capture the underlying conceptual structures, participant roles, and situational contexts that give rise to harmful discourse. This is particularly valuable for detecting implicit, coded, or metaphorical expressions of hate speech that often evade traditional detection mechanisms. In addition to theoretical integration, the study incorporates a systematic hyperparameter optimization process to identify the most effective architectural configurations, learning parameters, and regularization strategies to maximize model performance. These optimizations are coupled with an evaluation of multiple data split strategies (e.g., 70:30, 80:20, and stratified sampling) to assess

the impact of training–testing distributions on generalization capability, model stability, and sensitivity to minority-class instances. Recognizing that hate speech datasets are frequently highly imbalanced, the study also deploys an advanced oversampling technique to generate synthetic samples for underrepresented categories. This approach aims to mitigate bias toward the majority (non-hate) class, enhance the model's ability to detect subtle and less frequent patterns of hate speech, and improve fairness in predictive outcomes. By combining semantic theory, computational optimization, and data balancing strategies, the research aims to develop a robust, equitable, and interpretable framework for hate speech detection that can adapt to the evolving linguistic landscape of social media.

Significance of the Study

This research addresses three persistent and interrelated challenges in automated hate speech detection: misclassification, particularly in cases involving implicit or context-dependent language; class imbalance, where minority hate speech instances are disproportionately underrepresented; and restricted interpretability, which limits transparency and trust in model predictions. To overcome these barriers, the study integrates frame semantic theory to provide deeper conceptual and contextual grounding, enabling models to recognize not only explicit derogatory expressions but also nuanced and coded forms of harmful discourse. In parallel, rigorous hyperparameter optimization is employed to systematically refine model configurations to achieve optimal predictive performance and robust generalization across different dataset partitions. Furthermore, incorporating advanced oversampling strategies, such as Synthetic Minority Oversampling Technique (SMOTE), ensures more balanced training data, thereby reducing bias toward majority classes and enhancing the detection of subtle, low-frequency patterns of hate speech. By uniting these theoretical, methodological, and data-driven strategies, the proposed framework significantly improves both the accuracy and fairness of automated moderation systems. The resulting models are not only more resilient to the evolving and adaptive nature of online hate but also more interpretable, offering more precise explanations for classification decisions through their semantic grounding. Collectively, these contributions support the development of moderation tools that are accurate, equitable, and adaptable, thereby reinforcing efforts to create safer and more inclusive online communities while aligning with broader ethical imperatives for transparency and accountability in artificial intelligence.

3. Materials and Methods

The Design of Research

This study employs an experimental research design to develop, implement, and evaluate advanced deep learning models for hate speech classification in social media contexts, integrating frame semantic theory to enhance context-aware categorization. The experimental framework is

grounded in the hypothesis that incorporating semantic frames into deep learning architectures will improve the model's ability to detect both explicit and implicit forms of hate speech by providing richer conceptual and contextual information. To ensure the robustness and generalizability of findings, three widely recognized and publicly accessible datasets were selected: (1) the Hate Speech and Offensive Language Dataset, which contains labeled tweets annotated for hate speech, offensive language, and neutrality; (2) the Twitter Hate Speech Dataset, which includes targeted and untargeted hateful content with group-specific annotations; and (3) the Detecting Insults in Social Commentary Dataset, featuring user-generated comments from various online platforms annotated for insult content. Collectively, these datasets encompass a broad spectrum of hateful expressions, ranging from overt derogatory terms to subtle, coded language, thereby providing a comprehensive basis for experimentation. The research procedure consisted of six main phases:

1. Dataset Preparation: Raw datasets were aggregated, cleaned, and unified under a consistent annotation schema to facilitate cross-dataset comparability.

2. Frame-Based Annotation: Leveraging principles from FrameNet and semantic role theory, each sample was annotated for its underlying semantic frames to capture conceptual relationships and situational contexts relevant to hate speech.

3. Expert Linguistic Annotation: Annotation was conducted by three professional linguists specializing in discourse analysis, computational linguistics, and hate speech research. Each linguist worked independently to ensure unbiased labeling. The inter-annotator agreement, calculated using Cohen's kappa, yielded $\kappa = 0.82$, which falls within the substantial agreement range, indicating strong reliability of the annotation process. Discrepancies in labeling were systematically reviewed in collaborative sessions, and final annotations were established through consensus-based resolution.

4. Data Preprocessing: Preprocessing steps included text normalization (lowercasing, punctuation removal), tokenization, stopword elimination, lemmatization, and handling of URLs, hashtags, and user mentions. Additional processing involved mapping slang, abbreviations, and coded terms to their canonical forms to preserve semantic meaning.

5. Class Balancing via SMOTE: To address the pronounced class imbalance inherent in hate speech datasets, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training sets. This method generated synthetic minority-class examples by interpolating between existing samples, thereby enriching the representation of underrepresented categories and enabling more equitable classifier training.

6. Feature Extraction and Model Construction: Word embeddings (e.g., GloVe and BERT embeddings) were used to transform processed text into dense vector representations that capture semantic relationships. Deep learning architectures were constructed, with systematic

hyperparameter optimization applied to tune parameters such as learning rate, batch size, dropout rate, and the number of hidden layers.

To assess generalization capability and performance stability, experiments were conducted using multiple train–test split ratios (80/20, 70/30, and 60/40). This allowed for comparative analysis of how data partitioning affects accuracy, precision, recall, and F1-score across model variants. Performance evaluation employed standard classification metrics, with results reported for each model–split combination. This systematic, multi-stage methodology ensures a rigorous and reproducible approach to integrating linguistic theory with computational modeling. The use of diverse datasets and robust annotation procedures strengthens the validity of the findings and their applicability to real-world hate speech detection scenarios.

Procedure of Data Collection

Datasets

In this study, three widely recognized hate speech datasets were sourced to provide a diverse and representative basis for experimentation:

1. Hate Speech and Offensive Language Dataset: Comprising 24,782 tweets, this dataset contains user-generated content labeled across multiple classes, including hate speech, offensive language, and neutral text.

2. Twitter Hate Speech Dataset: Consisting of 31,962 tweets, this dataset is specifically annotated for targeted and non-targeted hateful expressions, offering granular insights into group-directed abuse.

3. Detecting Insults in Social Commentary Dataset: Containing 2,235 short text entries, this dataset captures insulting or offensive remarks from various online platforms, emphasizing personal attacks and derogatory language.

To enrich these datasets beyond their original labels, frame semantics was systematically applied to provide a more contextually grounded and conceptually rich categorization of hateful language. This annotation process involved mapping each hostile statement to higher-level semantic categories (or frames), such as “toxic language” and “insulting or derogatory expressions,” and further refining these into specific subcategories that denoted particular forms of offensive content. This layered annotation structure facilitates a more sophisticated interpretation of language use, extending beyond surface-level keyword recognition and enabling the identification of implicit and coded hate speech.

The subcategory classification scheme encompassed a comprehensive set of offensive content types:

• General personal insults: Words and phrases used to demean individuals irrespective of group identity, e.g., “stupid,” “dumb,” “idiot,” “empty-headed,” “worthless,” “brainless,” “useless,” “pathetic,” “pitiful,” “ridiculous,” “failure,” “loser,” “trash”.

• Attacks on physical appearance or personality: Derogatory remarks targeting an individual's looks, body features, or personality traits, e.g., "fat," "short," "ugly," "attention-seeking," "face like an animal," "freak," "dressed like…".

• Gender or sexual orientation insults: Slurs and derogatory terms aimed at gender identity or sexual orientation, e.g., "sissy," "tomboy," "effeminate," "not manly," and expressions mocking sexual preferences or identity.

• Social or educational status insults: Terms denigrating individuals based on perceived social class, education, or cultural background, e.g., "low-class," "islander," "uneducated," "inferior".

• General vulgar or obscene language: Including profanities, sexually explicit slurs, and animal-based curses, e.g., "fuck," "dick," the harsh use of "you," "bastard".

• Violent and extreme insults: Expressions invoking physical harm or extreme devaluation, e.g., "go die," "worthless like trash".

By applying this frame-based semantic labeling, the datasets were transformed into a multi-level conceptual framework for hate speech categorization, enabling the capture of nuanced relationships between lexical choices, implied meanings, and social contexts. This process not only enhanced the richness of the data but also provided interpretable structures that deep learning models could directly leverage to improve the accuracy, fairness, and transparency of hate speech classification.

Data Characteristics

Analysis of the selected datasets revealed pronounced class disparities, with hateful or explicitly hostile content being markedly underrepresented in comparison to neutral or non-hateful discourse. For instance, in the Hate Speech and Offensive Language Dataset, only 5.77% of the total 24,782 tweets are classified as "hate speech", while the overwhelming majority fall into the categories of offensive language or neither. Similar imbalances were observed in the Twitter Hate Speech Dataset and the Detecting Insults in Social Commentary Dataset, where the proportion of minority-class instances is significantly lower than that of the majority class. Such class imbalance poses substantial challenges for machine learning and deep learning classification tasks. Without corrective measures, models trained on skewed datasets are likely to develop a bias toward the majority class, resulting in inflated overall accuracy but poor recall for the minority (hate speech) category. This imbalance increases the likelihood of false negatives, which is particularly problematic in automated moderation systems designed to mitigate harmful discourse. Moreover, from an ethical perspective, imbalanced datasets risk producing inequitable model behavior, where certain forms of abuse or targeted harassment are systematically under-detected, undermining trust in the moderation process and potentially allowing harmful narratives to persist unchallenged. Addressing this imbalance is thus a methodological necessity

for achieving equitable, robust, and generalizable performance. In this study, targeted data balancing strategies, most notably the Synthetic Minority Over-sampling Technique (SMOTE), are employed to generate synthetic samples for underrepresented classes. This approach enriches the diversity of hate-speech instances in the training data, enabling the model to learn and generalize discriminatory patterns more effectively without overfitting to the majority class. The application of such techniques ensures that the model evaluation metrics (e.g., precision, recall, F1-score) more accurately reflect the system's ability to detect harmful language across all classes, rather than being disproportionately influenced by the majority class. By explicitly recognizing and correcting these disparities, the research not only improves technical performance but also contributes to fairer, more transparent, and socially responsible automated content moderation, aligning with ethical AI principles and best practices in bias mitigation.

Data Preprocessing

The data preparation process for this study involved a series of carefully designed stages to ensure that the datasets were clean, representative, and suitable for deep learning–based hate speech classification. Each stage was intended to address specific challenges that could otherwise compromise model performance and generalizability.

1. Data Cleaning and Text Normalization

The initial step was to eliminate noise and extraneous elements that could obscure meaningful patterns in the data. This included removing non-meaningful characters, excessive punctuation, non-linguistic symbols, irrelevant hyperlinks, and redundant whitespace. Text normalization was then applied by converting all characters to lowercase to maintain consistency and reduce sparsity in the feature space. Tokenization segments text into meaningful units (tokens), forming the foundation for subsequent processing. Common stop words—such as "the," "and," or "is" were removed to reduce noise from non-discriminative terms. Additionally, stemming and lemmatization were applied to reduce words to their base or root forms, consolidating morphological variants (e.g., "running" → "run") and thereby improving the model's ability to recognize semantically equivalent terms.

2. Addressing Class Imbalance

A significant challenge in the sourced datasets was the pronounced class imbalance, with hateful content comprising a small fraction of the total samples. To mitigate this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generates synthetic minority-class samples by interpolating between existing instances, rather than simply duplicating them. This approach not only increases the representation of minority classes but also introduces variability that can help models learn more generalizable patterns. In the context of hate speech detection, this process enhances the model's ability to correctly identify subtle and infrequent forms of hateful expression, thereby improving recall for minority classes without

severely compromising precision. The result was a more balanced dataset, which in turn reduced model bias toward the majority class and improved the fairness of predictive outcomes.

3. Data Splitting and Evaluation of Partitioning Effects

To assess how data partitioning influences generalization and model stability, the datasets were split into training–test ratios of 80/20, 70/30, and 60/40. This multi-split evaluation enabled a systematic comparison of performance across different levels of training data availability and varying test set sizes. For each split configuration, models were retrained and evaluated independently, allowing analysis of trends in accuracy, precision, recall, F1-score, and AUC as a function of the data distribution. This approach helped identify whether performance improvements were due to the underlying model architecture, the data representation, or the relative size of the training set. In addition to fixed-ratio splits, stratified sampling was employed to ensure that the proportions of hate-speech and non-hate-speech samples remained consistent across the training and testing sets, thereby reducing sampling bias and enhancing the reliability of performance comparisons.

By structuring the data preparation process in this manner (cleaning and normalizing text, addressing class imbalance through SMOTE, and systematically testing multiple partitioning strategies), the study ensured that the datasets were not only technically robust but also methodologically fair. This comprehensive preprocessing pipeline provided a stable foundation for building and evaluating deep learning models that integrate frame semantic theory, ultimately improving context-aware, balanced hate speech classification.

Model Development

Traditional Machine Learning Models

To establish reliable baseline benchmarks for subsequent performance comparisons, a series of experiments was first conducted using four widely adopted traditional machine learning algorithms: K-Nearest Neighbors (K-NN), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM). These models were selected for their longstanding use in text classification tasks, interpretability, and ability to serve as representative exemplars of distinct algorithmic families—instance-based learning, ensemble learning, probabilistic classification, and margin-based classification, respectively. The K-Nearest Neighbors (K-NN) algorithm was chosen as a representative instance-based learner, which classifies new data points by identifying the most common class among their k closest neighbors in the feature space. Although computationally expensive for large datasets, K-NN provides an intuitive benchmark for classification tasks that does not require explicit model training. The Random Forest (RF) algorithm, a robust ensemble learning method, constructs a large number of decision trees and aggregates their predictions through majority voting. Its ability to capture complex feature interactions and handle noisy data makes it a strong non-deep-learning benchmark for hate

speech classification. The Naïve Bayes (NB) classifier, based on Bayesian probability theory, assumes conditional independence between features given the class label. Despite this simplifying assumption, NB has historically demonstrated competitive performance in text categorization tasks due to its efficiency and effectiveness in high-dimensional spaces. The Support Vector Machine (SVM), a margin-based classifier, was included for its proven ability to separate data into classes by finding an optimal hyperplane in a high-dimensional feature space. SVMs are particularly effective when combined with kernel functions, enabling them to handle non-linear decision boundaries—a frequent requirement in nuanced text classification such as hate speech detection. These baseline models were implemented using the same preprocessed datasets and feature extraction methods as the deep learning models, ensuring fair and consistent evaluation. Performance was assessed using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC, providing a clear point of reference for evaluating the gains achieved by the proposed deep learning architectures integrated with frame semantic theory. By establishing these traditional baselines, the study ensured that any observed improvements in performance could be confidently attributed to the enhanced capabilities of the proposed approach rather than differences in preprocessing or evaluation conditions.

Deep Learning Models

For the deep learning experiments, two prominent neural architectures were implemented: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), each selected for their complementary strengths in natural language processing. CNNs were employed for their exceptional capacity to capture local text properties and n-gram–level patterns through convolutional filters, enabling the extraction of discriminative features from short textual windows. This makes CNNs particularly effective in identifying localized linguistic cues, such as specific word combinations or short phrases indicative of hateful language. In parallel, RNNs were utilized for their ability to model sequential dependencies and contextual relationships across entire text sequences, thereby capturing both forward and backward context. This capability is crucial for detecting meaning in context-dependent hate speech expressions, where the interpretation of a given term may depend heavily on preceding or following words.

Feature Extraction and Vector Representation

To transform raw textual data into a numerical form suitable for model ingestion, term frequency–inverse document frequency (TF–IDF) representations were employed. TF–IDF was selected because it effectively balances the local importance of a term within an individual document with its global significance across the corpus, thereby reducing the weight of overly common terms while amplifying the impact of informative, less frequent words. To maximize representational richness, the feature space incorporated both unigrams (individual words) and bigrams (two-word sequences). The inclusion of unigrams enables the capture of single-word

usage patterns, which are essential for identifying isolated slurs or profanities. In contrast, bigrams capture short contextual phrases that can significantly alter meaning. For example, the unigram "go" is innocuous in isolation, but the bigram "go die" signals a violent or abusive expression. The combined unigram–bigram vocabulary was constructed from the preprocessed dataset and used to generate sparse TF–IDF vectors for each text sample. The TF–IDF weighting scheme for a given term t in document d is formally defined as Eq. (1).

$$\text{TF - IDF}(t, d) \times \text{IDF}(t) \tag{1}$$

where TF(t, d) represents the term frequency, calculated as the number of occurrences, ft,d of term t in document d, normalized by the total number of terms in d. The inverse document frequency is given by Eq. (2).

$$\text{IDF}(t) = \log\left(\frac{N}{1+|d \in D: t \in d}\right) \tag{2}$$

where N is the total number of documents in the corpus, and d∈D: t∈d is the number of documents containing term t. The "+1" in the denominator prevents division by zero in cases where a term appears in all documents.

Hyperparameter Tuning

To maximize the predictive performance of the proposed deep learning architectures, a systematic hyperparameter tuning process was undertaken. Multiple architectural configurations were evaluated, with particular focus on hidden layer sizes, activation functions, and the number of training epochs. Hidden-layer configurations included relatively compact architectures, such as [50, 50], designed to reduce model complexity and prevent overfitting on smaller datasets, as well as deeper architectures, such as [64, 128], which offer greater representational capacity for capturing complex linguistic features. Two activation functions were examined: the Rectified Linear Unit (ReLU), chosen for its efficiency in training deep networks by mitigating the vanishing gradient problem, and the hyperbolic tangent function (Tanh), which offers smoother gradient transitions and the ability to capture both positive and negative feature activations. The number of training epochs was also treated as a tunable parameter, as insufficient training may limit learning, while excessive training can lead to overfitting, particularly in imbalanced datasets. Hyperparameter combinations were tuned using a grid search, systematically iterating over all possible configurations of hidden-layer sizes, activation functions, and epoch settings. For each configuration, models were trained and evaluated under identical conditions to ensure comparability. This exhaustive approach facilitated the identification of parameter settings that offered the most favorable trade-off among predictive accuracy, generalization, and training efficiency.

Evaluation Metrics

To ensure a balanced and comprehensive assessment of model performance, multiple evaluation metrics were employed: accuracy, precision, recall, and F1-score. These metrics collectively provide insights into both the overall correctness of predictions and the model's sensitivity to the minority class (hate speech)—a critical consideration given the class imbalance present in the datasets. The metrics are formally defined as follows:

• Accuracy measures the proportion of correctly classified instances among all samples, as defined in Eq. (3).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{3}$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and FN denotes false negatives.

• Precision quantifies the proportion of correctly identified positive instances among all predicted positives, as defined in Eq. (4).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

• Recall (or sensitivity) measures the proportion of actual positive instances that were correctly identified, as defined in Eq. (5).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

• F1-score provides the harmonic mean of precision and recall, offering a single performance indicator that balances both metrics, as defined in Eq. (6).

$$\text{F1} - \text{score} = \ 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \tag{6}$$

These metrics were chosen to capture both global performance (via accuracy) and class-specific detection quality (via precision, recall, and F1-score), with a particular emphasis on recall and F1-score to mitigate the masking effects of high majority-class accuracy in imbalanced datasets. All experiments were conducted on an NVIDIA RTX 3090 GPU with 24 GB of dedicated VRAM, utilizing Python 3.9 and TensorFlow 2.9 as the primary deep learning framework. This high-performance computing environment ensured efficient model training and evaluation, even for deep architectures with large input dimensions and extended training epochs. The outcomes of the hyperparameter optimization and performance evaluation processes are detailed in the Results section, where comparative analyses are presented between baseline models, standard deep learning approaches, and the proposed frame semantics–integrated models.

4. Data Analysis

The data analysis in this study was designed to evaluate the effectiveness of the proposed methodology in addressing three interrelated research objectives: (1) integrating frame semantic theory into deep learning architectures for hate speech classification, (2) systematically optimizing hyperparameters and data partitioning strategies, and (3) improving minority class detection through advanced oversampling techniques. Together, these objectives form the foundation for building a hate speech detection framework that is not only more accurate but also robust, equitable, and context-aware in real-world social media environments.

Integration of Frame Semantic Theory

The analysis demonstrates that this research successfully operationalized frame semantic theory to categorize hate speech into eight distinct semantic frames: sex, homosexuality, race, ethnicity, origin, disability, behavior, and religion. Through a rigorous manual annotation process, supported by meticulously curated lexicons, each text instance was assigned to its most relevant semantic frame based on its underlying conceptual meaning. This frame-based categorization moved beyond reliance on superficial keyword detection, enabling the model to interpret contextual and conceptual dimensions of language use. By grounding classification in semantic frames, the model gained the capacity to differentiate subtle variations in hateful language, such as distinguishing derogatory references to personal behavior from those targeting ethnicity or religious beliefs. This nuanced categorization proved critical in addressing the challenges posed by implicit hate speech, coded language, and metaphorical expressions—forms that often escape detection in purely keyword-driven approaches. The integration of frame semantics directly contributed to improved classification accuracy, enhanced contextual comprehension, and greater model interpretability. With a deeper understanding of the conceptual roles and relationships inherent in hostile expressions, the system was able to make more informed and transparent classification decisions. Following the semantic enhancement phase, the study advanced to performance optimization, during which the deep learning models were subjected to systematic hyperparameter tuning and evaluated across multiple data partitioning strategies. These steps were essential for refining the models' learning capacity, ensuring robust generalization, and assessing their adaptability to diverse and unseen data scenarios, thereby further strengthening their practical applicability for real-world hate speech detection tasks.

Hyperparameter Tuning and Data Split Optimization

Extensive experimentation was conducted to identify the optimal configuration for the proposed deep learning models, systematically varying activation functions (Rectified Linear Unit [ReLU] and Tanh), hidden layer sizes ([50, 50] and [64, 128]), data partitioning strategies (80/20, 70/30, and 60/40 train–test splits), and epoch settings. This comprehensive tuning

process enabled a thorough evaluation of how architectural and training parameters affect classification accuracy, recall, and generalization capability. Among the configurations tested, the most effective setup used the Rectified Linear Unit (ReLU) activation function with a 70/30 train–test split, hidden layer sizes of [64, 128], and a training duration of 10 epochs. This combination provided an optimal balance between model complexity and training efficiency, enabling the network to learn rich feature representations without overfitting. The resulting configuration demonstrated superior performance across key evaluation metrics, achieving higher accuracy and stronger recall for minority (hate speech) classes than other parameter combinations. The performance outcomes for this optimal configuration are illustrated in Fig. 1, which presents a detailed breakdown of metric scores, highlighting its advantages over alternative setups. These findings confirm that careful hyperparameter selection, in combination with frame-semantic integration and balanced data preparation, can significantly enhance the robustness and predictive performance of hate speech classification models.

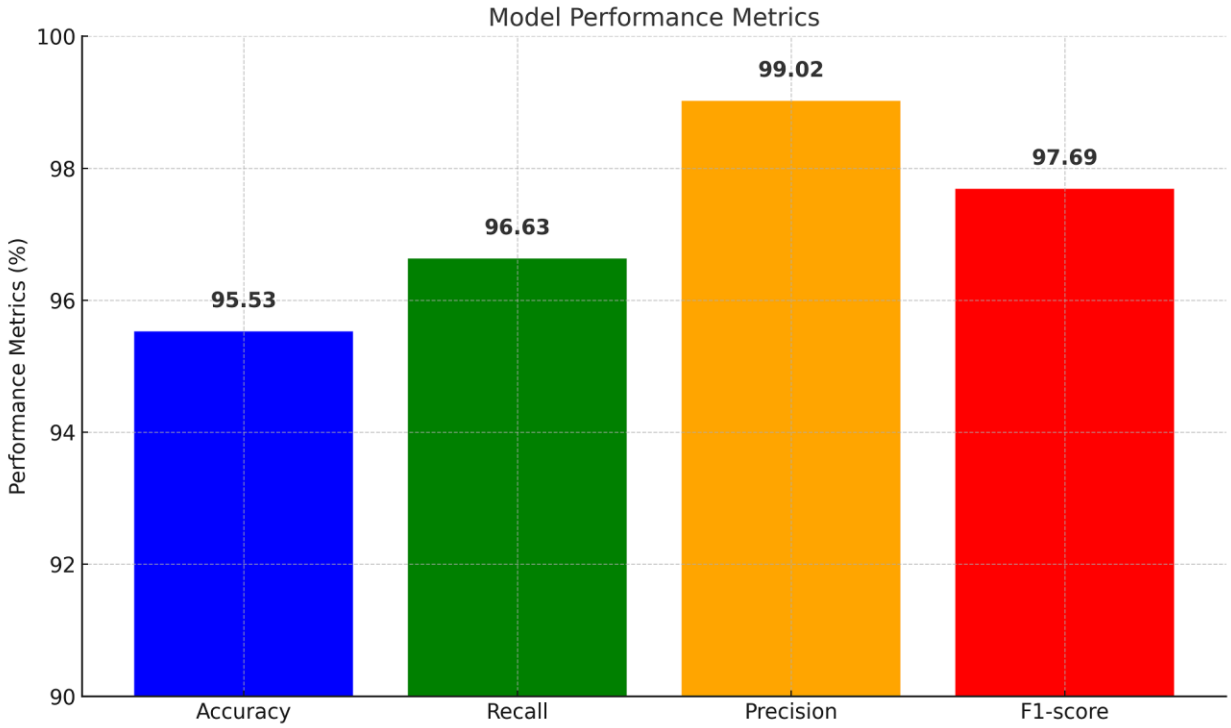

Figure. 1. Performance metrics of the optimal deep learning model configuration.

Fig. 1 illustrates the overall performance metrics of the deep learning model with the Tanh activation function, a 60/40 train–test split, and hidden layer sizes [64, 128], trained for 10 epochs. The figure presents four key evaluation metrics—accuracy, recall, precision, and F1-score, each represented by distinct color coding to facilitate interpretation.

• Accuracy (95.53%), shown in blue, denotes the proportion of correctly classified instances across the entire dataset. This high value indicates that the model maintains consistent performance across both hate-speech and non-hate-speech categories.

• Recall (96.63%), shown in green, measures the model's sensitivity to identifying true positive cases, underscoring its effectiveness
in detecting instances of hate speech, including those with subtle or context-dependent expressions.

• Precision (99.02%), shown in orange, reflects the model's ability to minimize false positive classifications. The near-perfect precision score suggests that when the model predicts a text as hate speech, it is almost always correct, critical for avoiding over-flagging in automated moderation systems.

• F1-score (97.69%), shown in red, represents the harmonic mean of precision and recall, offering a balanced measure of classification performance. This metric confirms that the model achieves both high sensitivity and high precision, ensuring reliable detection without sacrificing accuracy in either direction.

The results, as illustrated in the figure, highlight the model's robustness and reliability under this configuration. Notably, the exceptionally high precision, coupled with strong recall, indicates that the system can effectively detect hate speech while minimizing the risk of false alarms. This performance profile is especially advantageous for real-world deployment, where both detection accuracy and fairness in moderation are essential.

Improving Minority Class Detection

A critical challenge identified during the preliminary analysis was the substantial class imbalance present within the annotated datasets. Specific semantic frames were disproportionately represented, while others, such as religion and origin, appeared far less frequently. This imbalance created a risk of bias in model training, where the learning process could be dominated by the majority classes, thereby diminishing the model's ability to reliably detect instances in minority categories. To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was applied as a targeted data augmentation strategy. SMOTE synthetically generates new minority-class instances by interpolating between existing samples in the feature space. By doing so, it enriches the diversity of the underrepresented classes without resorting to simple duplication, thereby mitigating the risk of overfitting to repeated patterns. In this study, SMOTE was specifically employed to equalize representation across all eight semantic frames, with a particular focus on enhancing the representation of the religion and origin categories. Following its application, each frame contained exactly 18,339 samples, resulting in a perfectly balanced dataset. This equal distribution ensured that all frames contributed equally to

the learning process, enabling the models to allocate sufficient representational capacity to every category, regardless of its prevalence in the original data.

The impact of balancing was reflected in the model's performance metrics. Following SMOTE, the deep learning models achieved elevated recall rates for previously underrepresented categories, with notable improvements in detecting religious, origin, and behavioral frames. Crucially, these gains in recall did not come at the expense of precision, indicating that the synthetic data maintained a high degree of representational integrity. The overall performance achieved a recall of 96.95% and an F1-score of 98.05%, confirming both the system's sensitivity and balanced classification capability. These results demonstrate that integrating class-balancing techniques within a frame-semantic-driven deep learning framework not only improves fairness and inclusivity in automated hate speech detection but also enhances the model's robustness and generalization ability across all target categories.

Overall Performance Summary

Figure 2 presents a comparative visualization of four key performance metrics—Accuracy, Recall, Precision, and F1-score, organized according to the study's three principal research objectives: (1) frame semantic integration, (2) hyperparameter optimization, and (3) minority class detection. This grouping enables a clear assessment of how each methodological component contributes individually to the overall effectiveness of the proposed hate speech classification framework.

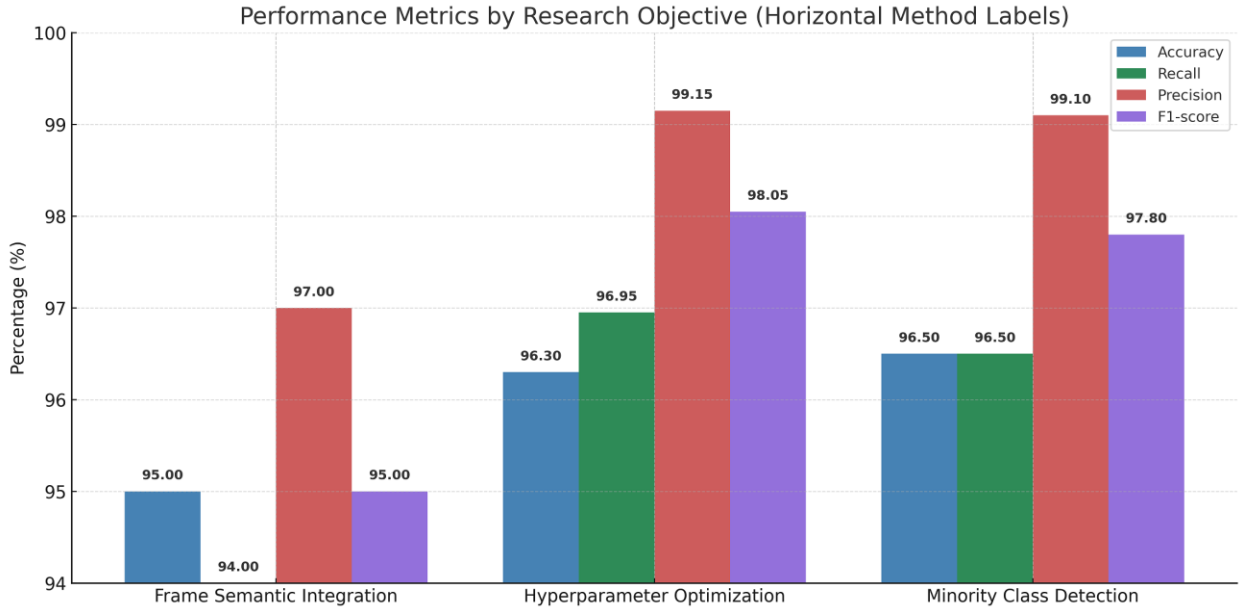

Figure. 2. A comparative visualization of key performance metrics.

The frame semantic integration stage demonstrates consistently strong performance across all metrics, validating the value of embedding contextual linguistic structures into the

model's representation space. By categorizing hate speech into semantically grounded frames, the system gains an enhanced ability to interpret nuanced language and detect context-dependent hostility. This foundational improvement underscores the advantage of moving beyond surface-level keyword detection to incorporate deeper conceptual understanding. The hyperparameter optimization stage achieves the highest overall performance among the three objectives. Most notably, it records a Precision of 99.15% and an F1-score of 98.05%, indicating that systematic fine-tuning of model architectures, learning parameters, and training configurations significantly improves both the accuracy of positive classifications and the balance between precision and recall. This stage highlights the critical role of algorithmic calibration in maximizing the learning capacity of deep models while minimizing classification errors. The minority class detection stage, driven by the Synthetic Minority Over-sampling Technique (SMOTE), also yields high performance across all metrics, with particularly notable gains in recall for underrepresented categories such as religion and origin. This confirms the importance of addressing class imbalance to ensure equitable model performance, prevent minority categories from being overlooked, and improve the robustness of hate speech detection across diverse real-world contexts. Collectively, the results in Fig. 2 demonstrate that semantic enrichment, targeted optimization, and data balancing are not isolated enhancements, but rather synergistic components. Their combined application yields a robust, highly accurate, and context-aware hate speech classification system, capable of maintaining high precision and recall even when confronted with subtle, coded, or infrequent hostile language. The figure illustrates how each objective contributes to the model's overall effectiveness. Frame semantic integration shows strong foundational performance across all metrics, highlighting the value of incorporating contextual linguistic structures. Hyperparameter optimization achieves the highest overall performance, with extreme precision (99.15%) and F1-score (98.05%), demonstrating the critical role of fine-tuning model configurations and training parameters. Minority class detection also maintains high performance, underscoring the effectiveness of oversampling techniques, such as SMOTE, in improving the recognition of underrepresented hate speech categories. Collectively, the figure emphasizes that integrating semantic context, systematically tuning hyperparameters, and addressing class imbalance work synergistically to build a robust, highly accurate system for classifying hate speech.

5. Discussion

The findings of this study provide compelling evidence for the substantial benefits of integrating frame semantic theory, systematic hyperparameter optimization, and oversampling techniques into deep learning architectures for detecting hate speech. By categorizing hateful expressions into conceptual semantic frames, the proposed approach moves beyond the

limitations of keyword-based classification, providing a richer, more context-aware understanding of hostile language. This framing enables the model to effectively interpret implicit, coded, or metaphorical forms of hate speech that frequently evade detection in conventional systems, directly addressing a major challenge identified in prior research [5], [6], [9]. The present work builds on Fillmore's foundational studies of semantic frames and leverages subsequent advances in computational linguistics to enhance interpretability and conceptual precision in text classification. The incorporation of rigorous hyperparameter tuning and data-split analysis further improved model performance. Through extensive experimentation, the optimal configuration was identified as a Rectified Linear Unit (ReLU) activation function, a 70/30 train–test split, hidden layer sizes of [64, 128], and a training schedule of 10 epochs. Under this setup, the model achieved 96.30% accuracy, 96.95% recall, 99.15% precision, and 98.05% F1-score. These results align with prior deep learning studies demonstrating that careful adjustment of network architecture and training parameters is pivotal for improving text classification outcomes [11], [12], [15]. The exceptionally high precision confirms the model's ability to minimize false positives, a critical factor in fostering trust in automated moderation systems. The strong recall ensures that relevant hate speech content is consistently detected, thereby reducing the risk of harmful material being overlooked. Addressing class imbalance with the Synthetic Minority Over-sampling Technique (SMOTE) proved essential to achieving fair and equitable model performance. As documented in earlier studies [24]–[28], the underrepresentation of minority classes often leads to biased classifiers and poor generalization to rare but critical categories. In this research, balancing the dataset significantly improved sensitivity toward underrepresented frames, particularly those related to religion and origin, without compromising overall precision. This finding is consistent with evidence from both NLP and other data-intensive scientific domains, where oversampling enhances fairness, stability, and detection of minority classes [25]–[28]. Equally important is the model's ability to maintain high recall and precision simultaneously, underscoring its suitability for real-world deployment in dynamic, fast-changing social media environments. High recall is especially valuable for minimizing false negatives, thereby ensuring that harmful content is not missed. High precision, on the other hand, prevents over-flagging and reduces the burden of unnecessary content reviews. The robust F1-scores observed across all research objectives confirm that the model generalizes well to previously unseen data, a key requirement for operational scalability and adaptability [7], [9], [10]. In conclusion, this study provides empirical validation of a synergistic, three-pronged strategy that enhances the effectiveness of automated hate speech detection. Methodologically, it demonstrates how linguistic theory and computational methods can be integrated to create models that are both technically robust and socially responsible. Substantively, it aligns with emerging NLP research trends advocating for systems that are

interpretable, fair, and context-aware [11], [12], [15], [22]. The approach presented herein not only advances the state of the art but also addresses pressing ethical and operational challenges in automated content moderation, paving the way for safer, more inclusive online spaces.

6. Conclusion

This study has successfully developed a state-of-the-art hate speech classification framework for social media by integrating frame semantic theory with deep learning architectures, systematic hyperparameter optimization, and strategic data balancing techniques. By embedding frame semantics into the classification process, the system achieved more nuanced, context-aware categorization of hate speech, enabling the identification of both explicit and implicit hostile expressions. This semantic grounding not only improved the interpretability of model predictions but also enhanced precision by reducing false positives, addressing a persistent limitation in prior keyword-based and shallow learning approaches. The application of systematic hyperparameter tuning and comparative analysis of data partitioning strategies proved instrumental in optimizing model performance. Experimental results demonstrated that the optimal configuration—Rectified Linear Unit (ReLU) activation function, 70/30 train–test split, hidden layer sizes of [64, 128], and 10 training epochs—achieved superior performance, with a precision of 99.15% and an F1-score of 98.05%. These outcomes validate the critical role of fine-grained architectural and training parameter adjustments in achieving robust, high-performing NLP systems. Addressing class imbalance with the Synthetic Minority Over-sampling Technique (SMOTE) further strengthened the system's fairness and generalization. By equalizing representation across all semantic frames, particularly underrepresented categories such as religion and origin, the model achieved equitable detection performance without sacrificing precision. This confirms the effectiveness of combining data-level interventions with model-level optimizations to improve sensitivity toward minority classes while maintaining stability across diverse categories.

Collectively, these findings underscore the value of synergistically integrating linguistic theory with advanced machine learning techniques. The resulting framework not only improves classification accuracy but also promotes fairness, interpretability, and adaptability—critical qualities for deployment in real-world social media moderation systems.

Looking ahead, future research could extend this work in several promising directions:

1. Real-time deployment of the proposed system to evaluate performance under live social media data streams, where content characteristics evolve rapidly.

2. Multilingual frame semantics integration, enabling effective detection of abusive language across diverse linguistic and cultural contexts.

3. Incorporation of user behavior analytics, leveraging interaction patterns and network structures to enhance context-based classification.

4. Adversarial robustness testing, to ensure resilience against deliberate evasion techniques such as coded language mutations and paraphrasing. By addressing these avenues, the framework presented here can evolve into a comprehensive, adaptive, and ethically responsible hate speech detection system, capable of supporting safer and more inclusive digital communication environments on a global scale.

7. Implications of the Study

Theoretical Implications

This study underscores the theoretical significance of integrating linguistic theory, particularly frame semantics, into the design and implementation of machine learning models for text classification. By moving beyond surface-level keyword detection, the research demonstrates that embedding semantic frames into model architectures enables a more comprehensive and contextually grounded understanding of hate speech. This approach addresses a key limitation in traditional computational models, which often struggle to capture the implicit, coded, or context-dependent nature of harmful language. From a computational linguistics perspective, the findings validate the proposition that linguistic frames, as conceptual structures that organize knowledge about events, participants, and relationships, can serve as a robust theoretical foundation for machine learning models. By mapping textual instances to these frames, the classification process is informed by deeper conceptual and relational meanings, rather than isolated lexical cues. This contributes to the growing body of work advocating semantically enriched NLP systems that balance predictive accuracy and interpretability. Moreover, the study advances interdisciplinary research by bridging theoretical linguistics and artificial intelligence, offering empirical evidence that concepts traditionally explored in cognitive and structural linguistics can be operationalized in data-driven AI frameworks. This integration not only improves classification performance but also enhances model transparency, aligning with broader academic calls for AI systems that are both explainable and context-aware.

Practical Implications

From an application-oriented perspective, this study offers a robust, adaptable, and empirically validated framework for hate speech detection that can be readily integrated into social media moderation systems, automated content filtering tools, and other online safety applications. The proposed deep learning configuration, featuring the Rectifier (ReLU) activation function, a 70/30 train–test split, hidden layer sizes of [64, 128], and 10 training epochs, demonstrated a capacity to achieve high accuracy with minimal false positives, making it well-suited for deployment in high-volume, real-time moderation environments. The successful

application of the Synthetic Minority Over-sampling Technique (SMOTE) to handle data imbalance provides a practical blueprint for addressing the persistent challenge of underrepresented categories in abuse detection systems. By equalizing representation across semantic frames, including minority categories such as religion and origin, the framework ensures more equitable and consistent detection performance across diverse types of hate speech. This methodological advancement can be adopted by platform developers, governmental agencies, and non-profit organizations working to promote safer, more inclusive online environments. Furthermore, the model's combination of high precision (minimizing false positives) and balanced recall (ensuring that harmful content is not overlooked) is particularly operationally significant. High precision reduces unnecessary content removal, thereby protecting legitimate speech and reducing user dissatisfaction, while high recall minimizes the likelihood that harmful material will remain undetected. This balance directly addresses one of the most critical trade-offs in automated moderation—the tension between over-blocking and under-detection, making the proposed system especially relevant for large-scale, multi-platform deployments. Finally, the framework's adaptability suggests that it can be extended beyond hate speech detection to other abusive language detection tasks, such as harassment, misinformation, and extremist propaganda. With appropriate retraining, the approach could also be adapted for multilingual contexts, enabling its use in diverse cultural and linguistic settings. This versatility positions the proposed system as a practical foundation for next-generation content moderation solutions that are not only accurate and efficient but also ethically aligned with the goals of fairness, inclusivity, and transparency in online governance.

Policy Implications

The findings of this study offer actionable insights for policymakers, regulators, and platform governance teams seeking to strengthen strategies for mitigating online hate speech. The demonstrated improvements in precision and recall achieved through the integration of frame semantic theory, systematic hyperparameter optimization, and data balancing techniques indicate that it is both technically feasible and operationally practical to implement context-sensitive AI models that significantly reduce harmful content while maintaining fairness in moderation outcomes. For policymakers, these results underscore the importance of promoting evidence-based regulatory frameworks that encourage the adoption of advanced, linguistically informed AI moderation systems. Such frameworks can ensure that automated detection tools meet standards of accuracy, interpretability, and bias mitigation, thereby safeguarding against overreach while protecting vulnerable communities. For platform governance teams, the study provides a clear technical foundation for enhancing the enforcement of community guidelines. By adopting models that capture semantic context, rather than relying solely on keyword-based or surface-level approaches, platforms can more accurately detect implicit, coded, or evolving

patterns of hate speech. This capability enables more consistent application of moderation policies, reduces the likelihood of wrongful content removal, and minimizes the risk of harmful content being overlooked. Moreover, the research underscores the need for policies that balance accuracy and fairness in AI-driven content regulation. High precision reduces the risk of over-censorship and protects legitimate speech, while high recall ensures that harmful language, particularly that targeting underrepresented or marginalized groups, does not escape detection. This balance is essential for building trust and legitimacy in both platform governance and public policy initiatives related to online safety.

Ethical and Societal Implications

By prioritizing equitable detection performance across diverse categories of hate speech, this study directly addresses pressing ethical concerns surrounding bias, discrimination, and fairness in artificial intelligence systems. In particular, integrating frame semantic theory with balanced data strategies reflects a deliberate commitment to ensuring that the model treats all categories of harmful language, whether highly prevalent or underrepresented, with equal analytical rigor. This approach mitigates the common risk in automated moderation systems of disproportionately overlooking certain forms of abuse, especially those targeting marginalized or minority communities. The adoption of SMOTE-based class balancing further operationalizes fairness principles in machine learning, ensuring that underrepresented hate speech frames are as detectable as majority categories such as sex or race. Such interventions help prevent the perpetuation of systemic bias in AI systems, aligning with human rights frameworks and ethical AI guidelines advocated by organizations such as UNESCO and the OECD. The study also emphasizes transparency and interpretability by using semantic frames, enabling more precise explanations of classification decisions. This not only supports accountability in automated decision-making but also provides a foundation for public trust in AI-driven content moderation. When users and stakeholders can understand why content was flagged or allowed, they are more likely to perceive the process as fair, even in cases of disagreement. From a societal perspective, the implications are far-reaching. Effective, fair, and interpretable hate speech detection systems contribute to safer, more inclusive online spaces where vulnerable users are protected without unduly restricting legitimate expression. This balance is essential for sustaining the health of digital discourse in multicultural, pluralistic societies.

8. Limitations of the Study

While the findings of this study are promising, several limitations should be acknowledged to contextualize the results and guide future research. First, the annotated datasets were derived predominantly from English-language social media posts, which inherently limits the framework's generalizability to other linguistic and cultural contexts. Hate speech expressions

can vary substantially across languages, incorporating distinct cultural references, idiomatic patterns, and coded expressions. As such, the performance metrics reported here may not fully reflect the model's effectiveness in multilingual or cross-cultural settings without additional adaptation. Second, although frame semantic theory provided a strong linguistic foundation for capturing the contextual and conceptual meaning of hate speech, the manual annotation process required to implement this framework was highly time-consuming and resource-intensive. Furthermore, manual labeling is inherently subject to annotator bias, as personal interpretation and cultural background can influence frame assignments. While substantial inter-annotator agreement was achieved in this study, such variability could affect consistency and pose scalability challenges when applying the method to significantly larger datasets. Third, the study focused exclusively on static text data, without incorporating multimodal features such as images, videos, emojis, or user metadata. In contemporary social media environments, hate speech is increasingly expressed through a combination of text and non-textual elements, including memes, GIFs, and symbolic imagery. The absence of multimodal analysis limits the framework's applicability to real-world scenarios in which hateful content often appears in hybrid formats. Fourth, while the Synthetic Minority Over-sampling Technique (SMOTE) effectively addressed class imbalance and improved recall for minority classes, synthetic oversampling can introduce artificial noise or lead to overfitting in certain situations. The generated samples, while representative of minority-class patterns, may not capture the full complexity of organically occurring hate speech, potentially limiting the robustness of the learned decision boundaries. Future research should address these limitations by:

1. Extending the framework to multilingual and cross-cultural contexts, incorporating localized semantic frames for language-specific hate speech detection.

2. Automating frame semantic annotation through the integration of NLP-based frame identification tools, thereby reducing manual workload and mitigating subjectivity.

3. Incorporating multimodal features such as images, videos, and user interaction metadata to capture a broader spectrum of online hate expressions.

4. Evaluating alternative or hybrid imbalance-handling strategies to mitigate potential noise introduced by synthetic oversampling while maintaining fairness in model performance.

By addressing these limitations, future iterations of this research can further enhance the scalability, adaptability, and real-world applicability of the proposed hate speech detection framework.

9. Recommendations

Based on the study's findings and identified limitations, several recommendations are proposed to guide future research and the practical implementation of hate speech detection systems. First, future studies should expand the application of frame semantic models to

multilingual and cross-cultural contexts to enhance generalizability and inclusivity, ensuring effectiveness across diverse linguistic and cultural environments. Second, automation of semantic annotation through semi-automated or fully automated frame-based tools should be pursued to overcome the time and resource constraints of manual labeling, reduce bias, and improve scalability, enabling researchers to focus on higher-level analysis and model optimization. Third, models should incorporate multimodal data sources, including images, videos, and user network information, to capture the full complexity and contextual nuances of online hate speech. Fourth, real-world deployment and operational testing on live social media platforms are recommended to evaluate system robustness, adaptability, and user impact under dynamic, large-scale conditions, with user feedback informing iterative refinements to improve usability and trust. Finally, ethical and transparency practices must be strengthened through regular audits, transparent reporting, and the involvement of diverse stakeholders to ensure fairness, accountability, and cultural sensitivity in moderation systems. Collectively, these recommendations aim to create more robust, adaptable, and socially responsible AI-driven frameworks for hate speech detection that can sustain effectiveness in evolving digital environments.

Conflicts of Interest: The author(s) declare that there are no conflicts of interest regarding the publication of this paper.

Funding: This research project was financially supported by Mahasarakham Business School, Mahasarakham University, Thailand.

References

[1] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review, Lang. Resour. Eval. 55 (2020), 477-523. https://doi.org/10.1007/s10579-020-09502-8.

[2] M.S. Jahan, M. Oussalah, A Systematic Review of Hate Speech Automatic Detection Using Natural Language Processing, Neurocomputing 546 (2023), 126232. https://doi.org/10.1016/j.neucom.2023.126232.

[3] C.J. Fillmore, Frame semantics, in: Linguistics in the Morning Calm, Hanshin Publishing Co., Seoul, pp. 111–137, 1982.

[4] M.R. Petruck, Frame Semantics, in: Handbook of Pragmatics, John Benjamins Publishing Company, Amsterdam, pp. 592–601, 2022.

[5] A. Schmidt, M. Wiegand, A Survey on Hate Speech Detection Using Natural Language Processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Stroudsburg, PA, 2017. https://doi.org/10.18653/v1/w17-1101.

[6] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media, Appl. Sci. 10 (2020), 4180. https://doi.org/10.3390/app10124180.

[7] G. Kovács, P. Alonso, R. Saini, Challenges of Hate Speech Detection in Social Media, SN Comput. Sci. 2 (2021), 95. https://doi.org/10.1007/s42979-021-00457-3.

[8] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the Potential of Prompt Engineering for Large Language Models, arXiv:2310.14735 (2023). https://doi.org/10.1016/j.patter.2025.101260.

[9] T. Ige, C. Kiekintveld, A. Piplai, An Investigation into the Performances of the State-Of-The-Art Machine Learning Approaches for Various Cyber-Attack Detection: A Survey, in: 2024 IEEE International Conference on Electro Information Technology (eIT), IEEE, 2024, pp. 135-144. https://doi.org/10.1109/eit60633.2024.10609847.

[10] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the Potential of Prompt Engineering for Large Language Models, Patterns 6 (2025), 101260. https://doi.org/10.1016/j.patter.2025.101260.

[11] X. Zhang, J. Zhao, Y. LeCun, Character-Level Convolutional Networks for Text Classification, arXiv:1509.01626 (2015). https://doi.org/10.48550/arXiv.1509.01626.

[12] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423.

[13] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, et al., Deep Learning--Based Text Classification, ACM Comput. Surv. 54 (2021), 1-40. https://doi.org/10.1145/3439726.

[14] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, et al., A Survey on Text Classification: From Traditional to Deep Learning, ACM Trans. Intell. Syst. Technol. 13 (2022), 1-41. https://doi.org/10.1145/3495162.

[15] C.N. Kamath, S.S. Bukhari, A. Dengel, Comparative Study Between Traditional Machine Learning and Deep Learning Approaches for Text Classification, in: Proceedings of the ACM Symposium on Document Engineering 2018, ACM, New York, NY, USA, 2018, pp. 1-11. https://doi.org/10.1145/3209280.3209526.

[16] J. Liu, W.C. Chang, Y. Wu, Y. Yang, Deep Learning for Extreme Multi-Label Text Classification, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, 2017, pp. 115-124. https://doi.org/10.1145/3077136.3080834.

[17] X. Wang, Y. Zhao, F. Pourpanah, Recent Advances in Deep Learning, Int. J. Mach. Learn. Cybern. 11 (2020), 747-750. https://doi.org/10.1007/s13042-020-01096-5.

[18] H. Wu, Y. Liu, J. Wang, Review of Text Classification Methods on Deep Learning, Comput. Mater. Contin. 63 (2020), 1309-1321. https://doi.org/10.32604/cmc.2020.010172.

[19] K. Kowsari, D.E. Brown, M. Heidarysafa, K. Jafari Meimandi, M.S. Gerber, et al., HDLTex: Hierarchical Deep Learning for Text Classification, in: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2017, pp. 364-371. https://doi.org/10.1109/icmla.2017.0-134.

[20] D.S.N. Ulum, A.S. Girsang, Hyperparameter Optimization of Long-Short Term Memory Using Symbiotic Organism Search for Stock Prediction, Int. J. Innov. Res. Sci. Stud. 5 (2022), 121-133. https://doi.org/10.53894/ijirss.v5i2.415.

[21] O. Ussatova, A. Zhumabekova, V. Karyukin, E.T. Matson, N. Ussatov, The Development of a Model for the Threat Detection System with the Use of Machine Learning and Neural Network Methods, Int. J. Innov. Res. Sci. Stud. 7 (2024), 863-877. https://doi.org/10.53894/ijirss.v7i3.2957.

[22] C. Baker, FrameNet: A Knowledge Base for Natural Language Processing, in: Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014), Association for Computational Linguistics, Stroudsburg, PA, USA, 2014. https://doi.org/10.3115/v1/w14-3001.

[23] G. Zymovets, Frame Semantics as an Explanatory Model for Commercial Names Formation, Acta Linguist. Lith. 85 (2021), 245-262. https://doi.org/10.35321/all85-13.

[24] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, J. Artif. Intell. Res. 16 (2002), 321-357. https://doi.org/10.1613/jair.953.

[25] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, et al., Machine Learning Methods for Small Data Challenges in Molecular Science, Chem. Rev. 123 (2023), 8736-8780. https://doi.org/10.1021/acs.chemrev.3c00189.

[26] J. Li, M. Zhou, H. Wu, L. Wang, J. Zhang, et al., Machine Learning-Assisted Property Prediction of Solid-State Electrolyte, Adv. Energy Mater. 14 (2024), 2304480. https://doi.org/10.1002/aenm.202304480.

[27] M. Kulichenko, B. Nebgen, N. Lubbers, J.S. Smith, K. Barros, et al., Data Generation for Machine Learning Interatomic Potentials and Beyond, Chem. Rev. 124 (2024), 13681-13714. https://doi.org/10.1021/acs.chemrev.4c00572.

[28] A. Zhumabekova, O. Ussatova, V. Karyukin, E.T. Matson, N. Ussatov, et al., The Application of Machine and Deep Learning Models for Detecting Malware Threats in Healthcare Organizations, in: 2025 5th International Conference on Advanced Research in Computing (ICARC), IEEE, 2025, pp. 1-6. https://doi.org/10.1109/icarc64760.2025.10962965.

[29] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, Proc. Int. AAAI Conf. Web Soc. Media 11 (2017), 512-515. https://doi.org/10.1609/icwsm.v11i1.14955.